

**Uživatelsky rozšiřitelný slovník založený
na spolupráci davu
User-Extensible Dictionary a Crowd
Sourcing Project**

Zadání diplomové práce

Student:

Bc. Vojtěch Hložánka

Studijní program:

N2647 Informační a komunikační technologie

Studijní obor:

2612T025 Informatika a výpočetní technika

Téma:

**Uživatelsky rozšiřitelný slovník založený na spolupráci davu
User-Extensible Dictionary a Crowd Sourcing Project**

Jazyk vypracování:

čeština

Zásady pro vypracování:

Cílem práce je návrh a implementace slovníku pro češtinu a angličtinu založeného na spolupráci davu s webovou podporou.

1. Vytvořte webový server podporující uživatelsky přívětivou tvorbu slovníku založenou na principech spolupráce davu (crowd sourcing).
2. Navrhněte vhodné formáty dat a protokoly pro jejich sdílení, postupy pro validaci slovníkových pojmů za případné spolupráce experta.
3. Inspirujte se stávajícími přístupy a srovnajte s nimi výsledné řešení, realizované nejméně do fáze funkčního prototypu.

Seznam doporučené odborné literatury:

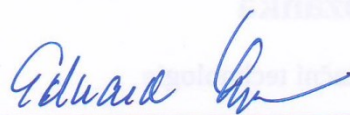
- [1] MLÝNKOVÁ, Irena, POKORNÝ, Jaroslav. XML technologie: principy a aplikace v praxi. 1. vyd. Praha: Grada, 2008, 267 s. Průvodce (Grada). ISBN 978-80-247-2725-7.
- [2] VIRIUS, Miroslav. C# 2010: hotová řešení. 1. vyd. Brno: Computer Press, 2012, 424 s. K okamžitému použití (Computer Press). ISBN 978-80-251-3730-7.
- [3] Atlantida Multilingual Dictionary [online]. 2006 [cit. 2014-03-31]. Dostupné z: <http://atla.revdanica.com/en/>
- [4] GOOGLE. Google translate [online]. 2007 [cit. 2014-03-31]. Dostupné z: <https://support.google.com/translate/#>
- [5] What is reCAPTCHA. ReCaptcha [online]. 2014 [cit. 2014-03-31]. Dostupné z: <http://www.google.com/recaptcha/learnmore>
- [6] Duolingo [online]. 2013 [cit. 2014-03-31]. Dostupné z: <https://www.duolingo.com/>
- [7] NĚMEC, Jiří. SlovníQ [online]. 2010 [cit. 2014-03-31]. Dostupné z: <http://code.google.com/p/slovníq/>
- [8] XDXF Description. Github [online]. 2013 [cit. 2014-04-01]. Dostupné z: https://github.com/soshial/xdxf_makedict/blob/master/format_standard/xdxf_description.md
- [9] PATTON, R. Testování softwaru. SAMS, Computer Press, Praha, 2002, ISBN 80-7226-636-5
- [10] PALETA, Petr. Co programátory ve škole neučí: aneb Softwarové inženýrství v reálné praxi. Vyd. 1. Brno: Computer Press, 2003, 337 s. ISBN 80-251-0073-1.

Formální náležitosti a rozsah diplomové práce stanoví pokyny pro vypracování zveřejněné na webových stránkách fakulty.

Vedoucí diplomové práce: **doc. RNDr. Petr Šaloun, Ph.D.**

Datum zadání: 01.09.2015

Datum odevzdání: 29.04.2016



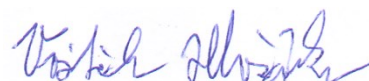
doc. Dr. Ing. Eduard Sojka
vedoucí katedry



prof. RNDr. Václav Snášel, CSc.
děkan fakulty

Prohlašuji, že jsem tuto diplomovou práci vypracoval samostatně. Uvedl jsem všechny literární prameny a publikace, ze kterých jsem čerpal.

V Hlučíně 20. 4. 2016


.....

Dovolte mi prosím, abych v úvodu své diplomové práce vyjádřil vřelé poděkování doc. RNDr. Petru Šalounovi Ph.D., Mgr. Aleně Kašpárkové, Ph.D. a Jaroslavu Pavláskovi.

Docent Petr Šaloun mi pomáhal již od roku 2013 při řešení mé bakalářské práce a nyní mě průběžně podporoval až k finální podobě této diplomové práce.

Magistra Alena Kašpárková mi laskavě věnovala svůj čas a obětavě poskytla zpětnou vazbu z profesionálního překladatelského hlediska.

Pan Jaroslav Pavlášek, dlouholetý člen operního sboru Národního Divadla Moravskoslezského v Ostravě, mi poskytoval cenné konzultace s ohledem na krásu českého jazyka.

Abstrakt

Tato diplomová práce navazuje na mou předchozí bakalářskou práci „Uživatelsky rozšiřitelný slovník“, kde byla vytvořena aplikace určena pro lokální počítač (tlustý klient). Tato aplikace nabízela uživateli standardní uživatelské rozhraní s možnostmi tvorby vlastních překladů a také jejich sdílení. Tato práce se zabývá sdílením uživatelských slovníků pro rozvíjení jejich rozsahu a funkcí, s možností ovlivnit kvalitu jejich obsahu. Tímto se využívá komunitní spolupráce (crowdsourcing). Pro zajištění správnosti překladů je vytvořena role správce. Webový server byl vytvořen v jazyce C# s technologií .NET. Data slovníku jsou uložena v standardu XDXF v XML, který umožňuje aktualizaci, případně získání obsahu slovníkových dat. Součástí je i anglicko-český slovník, který je možné rozšiřovat.

Klíčová slova: komunitní spolupráce; .NET; C#; Překladový slovník; XDXF; XML

Abstract

This master thesis builds on my previous bachelor thesis „User-Extensible Dictionary“, where there was created a local desktop application (fat client). The dictionary offered common user interface with the possibility to create user-specific translations and their sharing. By sharing user dictionaries, the scope and possibilities of the dictionary are expanded further, including the chance of improving the quality of its content. This sharing means crowdsourcing is involved. Administrator role was created because of the need for quality checks. Web server was created in C# language with .NET technology. Dictionary data is saved in XML using XDXF standard, which allows obtaining or updating the content of dictionary data. A basic English-Czech dictionary is included. This dictionary can be expanded further.

Keywords: crowdsourcing; .NET; C#; Translation dictionary; XDXF; XML

Seznam použitých zkratek a pojmů

crowdsourcing	komunitní spolupráce
game with purpose	hra s účelem
modifikovaný slovník	slovník obsahující spojené data uživatelů, kteří sdíleli svůj slovník
oficiální slovník (OS)	slovník daného překladu
tag	XML značka/tag, významově je uzlem XML stromu
uživatelský slovník	slovník obsahující data uživatele
CRUD	Create Read Update Delete; vytvoření, čtení, editace a smazání záznamu
GUI	Graphic User Interface; uživatelské grafické rozhraní
XDXF	XML Dictionary Exchange Format
XML	Extensible Markup Language; značkovací jazyk
XSD	XML Schema Definition; XML schéma

Obsah

1. Úvod	4
2. Současný stav	5
2.1. Překladač Google	5
2.1.1. Použití	6
2.2. Překladač Seznam	8
2.2.1. Použití	8
3. Crowdsourcing	10
3.1. Hry s účelem	10
3.1.1. EteRNA	10
3.1.2. Astro Drone [5]	11
3.2. Made	12
3.3. Duolingo	12
3.4. reCaptcha	13
4. XDXF	15
4.1. Úpravy pro uživatelskou rozšiřitelnost	15
5. Architektura & návrh	19
5.1. ASP.NET	20
5.1.1. Použité prvky	21
5.2. Role	21
5.2.1. Anonymní uživatel	21
5.2.2. Přihlášený uživatel	22
5.2.3. Správce	22
5.3. Kontrola kvality	22
5.4. Formální vytipování místa na trhu (cílová skupina)	23
5.5. Životní cyklus	24
5.5.1. Nahrání slovníku	24
5.5.2. Spojení slovníků	24
5.5.3. Ověření správcem	26
5.6. Komprimace souborů	28
6. Uživatelské rozhraní stránek	29
6.1. Master Page	29

6.2.	Stránky slovníků.....	30
6.3.	Stránka správce (Administrátora).....	30
7.	Implementace	32
7.1.	Překlad.....	32
7.2.	Slovník	33
7.3.	Nahrání slovníku a jeho aktualizace.....	34
7.4.	Stažení slovníku	36
7.5.	Kontrola slovníku správcem.....	37
7.5.1.	Odstranění redundancí.....	37
7.5.2.	Navigace v slovníku	38
7.5.3.	Přidání překladu	38
7.6.	Komprimace slovníku	40
8.	Testování	41
9.	Budoucí rozšíření	42
10.	Závěr.....	44
11.	Reference.....	45
Příloha A.	Obsah přiloženého DVD	47
Příloha B.	Formát XDXF.....	48

Seznam obrázků

OBRÁZEK 1 UKÁZKA VÝSLEDKŮ	7
OBRÁZEK 2 ZOBRAZENÍ DEFINICE	7
OBRÁZEK 3 PŘEKLAD SLOV	8
OBRÁZEK 4 HISTORIE	8
OBRÁZEK 5 PŘÍKLAD ÚKOLU ETERNA	11
OBRÁZEK 6 DOKOVÁNÍ NA ISS	11
OBRÁZEK 7 KONTROLA PŘEKLADU TEXTU	13
OBRÁZEK 8 ReCAPTCHA KONTROLA	13
OBRÁZEK 9 ReCAPTCHA ZKOMOLENÉ ZNAKY	13
OBRÁZEK 10 DIAGRAM KOMPONENT SERVERU	19
OBRÁZEK 11 DATA MAPPER	20
OBRÁZEK 12 FUNKČNOST ASP.NET[18]	20
OBRÁZEK 13 ARCHITEKTURA ASP.NET	21
OBRÁZEK 14 PŘÍPAD UŽITÍ, PRÁVA	22
OBRÁZEK 15 AKTUALIZACE SLOVNÍKU	23
OBRÁZEK 16 KOMUNIKACE SE SERVEREM PŘI SDÍLENÍ SLOVNÍKU	23
OBRÁZEK 17 NAHRÁNÍ SLOVNÍKU NA SERVER	24
OBRÁZEK 18 SPOJENÍ PŘEKLADŮ	25
OBRÁZEK 19 PŘIDEJ PŘEKLAD	27
OBRÁZEK 20 PŘIDEJ KLÍČOVÉ SLOVO	27
OBRÁZEK 21 VERTIKÁLNÍ MENU	29
OBRÁZEK 22 HORIZONTÁLNÍ MENU	29
OBRÁZEK 23 STRÁNKA SLOVNÍKŮ	30
OBRÁZEK 24 STRÁNKA SPRÁVCE	31
OBRÁZEK 25 DEFINICE	32
OBRÁZEK 26 TŘÍDA SLOVNÍK	34

1. Úvod

Profesionální slovníky v elektronické podobě jsou běžně komerčně i volně dostupné. Mnoho webových stránek umožňuje uživatelům aktivně se zapojit. To v sobě zahrnuje možnost spojit vzájemně své úsilí jak po stránce obsahu, tak ve sdílení výsledků. Tento trend může být vhodný pro rozšíření slovníků, které jsou uživateli vytvářeny společně – budou-li komunitou přijaty a rozvíjeny. Tímto rozšířením mohou být vytvořeny i slovníky odborné, jako například lékařský nebo účetní. Dobrým východiskem pro webový slovník jsou vhodně navržené tagy, spolu se strukturami a slovníkovými daty v XML, které je možno dále rozšiřovat.

Cílem mé Diplomové práce je vytvoření slovníku, který budou moci uživatelé rozvíjet. K tomu budou použity: Aplikace, kde uživatel vytváří svůj slovník, která pracuje bez potřeby aktivního internetového připojení (tuto aplikaci jsem vytvořil v mé bakalářské práci¹ [2]). Dále, webová aplikace, kterou jsem vytvořil jako součást Diplomové práce. Tyto práce jsou modifikovaný koncept klient-server, popřípadě distribuovaná textová databáze s definovaným způsobem sdílení. Uživatel sdílením svého slovníku a s hodnocením překladů, bude pomáhat k tvorbě přesnějšího slovníku.

V textové části, v 2. kapitole je popsán současný stav a slovníky, kterými jsem se inspiroval. Kapitola 3. popisuje crowdsourcing a jeho úspěšné příklady. Následně, ve 4. kapitole, je uveden formát XDXF, který umožňuje uživateli vytvářet a rozšiřovat slovník. Pak, v 5. kapitole, je navržena architektura a návrh funkčnosti aplikace. Kapitola 6. vysvětluje vzhled uživatelského rozhraní. Kapitola 7. popisuje implementaci aplikace. Postupně jsou objasněny funkce programu s jeho přednostmi i nedostatky. Kapitola 8. představuje testování a kontrolu aplikace předpokládanými uživateli i správci. Na závěr, v 9. kapitole, je prezentován návrh budoucího rozšíření programu. Při vytváření tohoto programu byla použita jako užitečný zdroj informací stránka StackOverflow² [1]. Tato práce využívá části textů mé bakalářské práce [2] .

¹ HLOŽÁNKA, Vojtěch *Uživatelsky rozšiřitelný slovník (User-Extensible Dictionary)*. Ostrava, Obhájeno 2014. Bakalářská práce. VŠB - Technická univerzita Ostrava, Fakulta elektrotechniky a informatiky, Katedra informatiky. Vedoucí práce doc. RNDr. Petr Šaloun Ph.D.

² Stackoverflow. C#. *Stackoverflow* [online]. 2013 [cit. 2016-03-25]. Dostupné z: <http://stackoverflow.com/questions/tagged/c%23>

2. Současný stav

Jak jsem se již zmínil v úvodu, tato práce navazuje na moji bakalářskou práci [2]. Zde byl vytvořen tlustý klient, pomocí kterého mohl uživatel používat CRUD operace nad svými slovníkovými daty. Uživatel měl možnost si vytvořit vlastní uživatelský slovník. Další možností bylo načtení oficiálního slovníku se stejným jazykovým překladem např. „ENG–CZE“, kde uživatel měl možnost upravovat překlady v oficiálním slovníku. Tyto úpravy se ukládaly do uživatelského slovníku.

Vizí je: vytvořený slovník sdílet s ostatními uživateli nahráním na server přes webové rozhraní. Po každém nahrání na server se tento uživatelský slovník strojově spojí s oficiálním. Aby byla možná kvalitativní kontrola slovníku, tak pro každý jazyk by měl existovat alespoň jeden správce, který by se o tento slovník staral. Při každém nahrání uživatelem, bude vytvářet dočasný (modifikovaný) slovník. Ten bude pouze pro zjednodušení práce správce. Tedy správce bude muset projít jeden slovník a ne kontrolovat postupně nahrané uživatelské slovníky. Při kontrole modifikovaného slovníku má správce možnost vložit celý překlad, případně klíčové slovo do oficiálního slovníku. Má také možnost použít pouze část překladu, jako například všechno, až na příklady. Nebo přidat samotné číselné ohodnocení překladu, kategorie a podobně.

Pro počáteční anglicko-český slovník je používán GNU/FDL Anglicko-Český slovník¹ [3], který byl z textového formátu převeden do XDXF. Obsahuje 87 485 různých překladů.

Analýza dostupných slovníků

V současnosti je k dispozici řada různých slovníků a to jak překladových, tak výkladových. Tyto slovníky, ať už se jedná o komerční či ne, nejsou uživatelsky rozšiřitelné. U běžně dostupných slovníků uživatel nemůže přidat ke slovu další překlad, natož aby mu bylo umožněno přidat nové slovo, a to včetně překladu. Jsou pouze pro čtení.

2.1. Překladač Google

Překladač Google je bezplatný automatický překladač, jehož provoz začal v roce 2007, spuštěný společností Google Inc. Překladová služba je zabudována v internetovém prohlížeči Google Chrome. Zde automaticky detekuje cizí jazyk a ve formě vysouvací nabídky pod adresním řádkem a nabízí okamžité přeložení stránky či dokumentu. Čeština byla do překladače Google² [9] přidána v květnu 2008. V současnosti překladač Google podporuje překlady desítek jazyků. Tento překladač má výhodu v tom, že je možné upřesňovat překlady. Tedy když se nám nelíbí překlad, je možné navrhnout nový. Toto upřesnění není okamžitě patrné, aplikuje se až tehdy, když stejné upřesnění navrhlo více uživatelů. Je tímto využíván crowdsourcing.

¹ GNU/FDL Anglicko-Český slovník [online]. [cit. 2016-03-31]. Dostupné z: <http://slovník.zcu.cz>

² GOOGLE. Google translate [online]. [cit. 2016-02-12]. Dostupné z: <https://support.google.com/translate/#>

2.1.1. Použití

Překladač Google má mnoho funkcí:

- okamžitý překlad namířením fotoaparátu
- změna klávesnice za účelem vyhledávání v jiných jazycích
- překlady:
 - fotky pořízené zařízením
 - obrázků uložených v zařízení
 - ručně psaného textu
 - webových stránek a dokumentů
 - offline
 - hlasového zadání

Zde je část podrobně popsána.

Okamžitý překlad textu na stránce

1. Výběr jazyku pro překlad. Když není jisté, v jakém jazyce je daná věta, lze použít funkci **Rozpoznat jazyk**. S vyšším množstvím zadaného textu se přesnost automatického rozpoznání jazyka zvyšuje.
2. Při psaní se překlad se začne ihned zobrazovat.

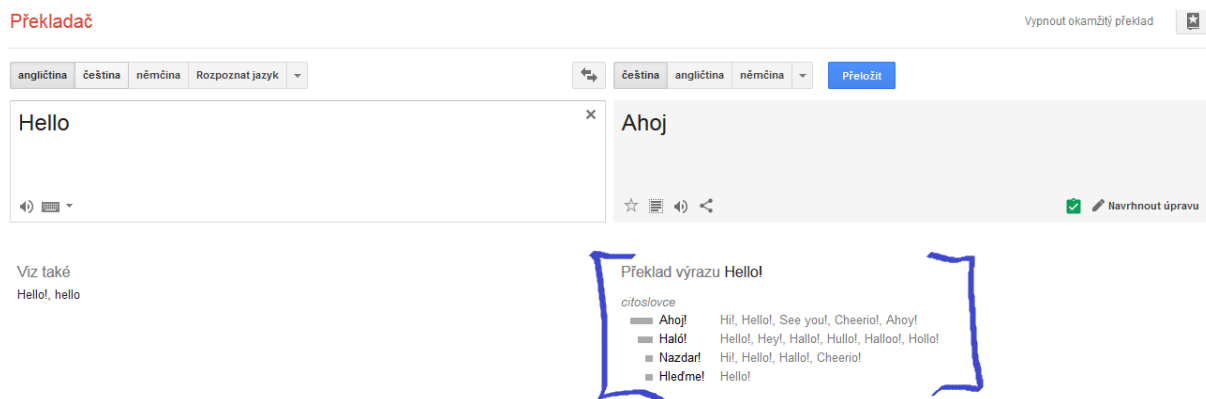
Čtení a poslech překladu

Pokud se překládá do jazyka, který nepoužívá latinku, zobrazí se u překladu tlačítko se symbolem **Ä**. Kliknutím na toto tlačítko se ukáže překlad přepsaný latinkou.

U mnoha jazyků je také možnost u přeloženého textu použít tlačítko s reproduktorem **🔊**. Když se na něj kliknete, přehraje se vyslovená verze překladu.

Výsledky ze slovníku

V případě, že se přeloží jedno slovo nebo běžná fráze, může se pod překladem zobrazit jednoduchý slovník se slovními druhy a možnými alternativními překlady (Obrázek 1). Vedle každého slovníkového hesla lze vidět odpovídající sadu zpětných překladů do původního jazyka. Velikost pruhu vedle každého hesla ukazuje relativní četnost výskytu překladu překlad na internetu.



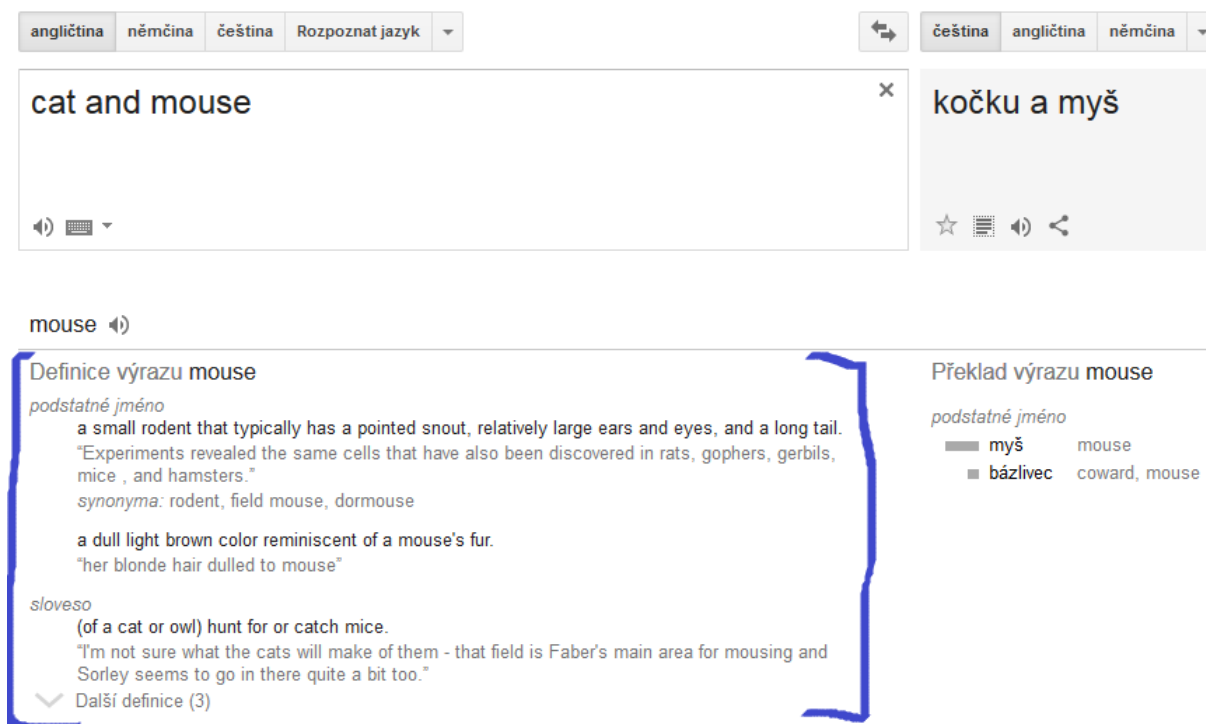
Obrázek 1 Ukázka výsledků

Ukládání překladů do slovníčku frází

Je možné vytvořit slovníček frází s překlady, které byly v minulosti hledány a které mohou být opět potřeba. Slovníček frází je přístupný offline a na webu translate.google.com.

Zobrazení definic použití slov

Pokud se hledají překlady jednotlivých slov nebo frází (Obrázek 2), je často užitečné vidět překlad v kontextu. Označením slova, se zobrazí definice slova. Slouží tedy jako výkladový slovník.



Obrázek 2 Zobrazení definice

2.2. Překladač Seznam

Slovník web portálu Seznam¹ [10] nemá sice tolik možností jako Google překladač, je však více rozšířen mezi neanglicky mluvícími Čechy a Slováky. Umožňuje překlad z/do 8 jazyků: anglicky, německy, francouzsky, italsky, španělsky, rusky, slovensky a samozřejmě česky. Je schopen, nejen překládat mezi jazyky, ale lze jej použít jako výkladový slovník.

2.2.1. Použití

Pro využití většiny možností, které tento překladač nabízí je potřeba přihlášení. Samotný překlad slovíček přihlášení nepožaduje.

Překlad

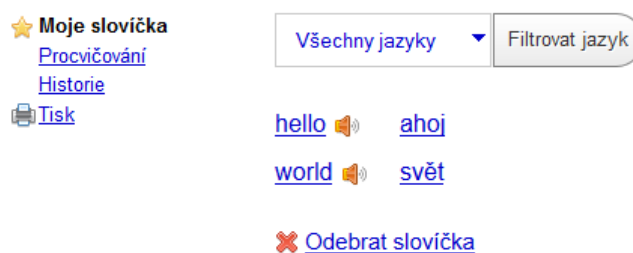
Slovník neumožňuje překlad celých vět. Namísto toho nabídne uživateli překlad jednotlivých slovíček (Obrázek 3). Pro napsání slov, které používají jiné znaky, než jsou na klávesnici, je možnost použít virtuální klávesnici, která simuluje standartní klávesnice jazyků, které tento překladač podporuje.



Obrázek 3 Překlad slov

Ukládání

Slovník umožňuje ukládání slovíček, které jsou vyhledány (Obrázek 4). Do „mých slovíček“ se vloží frázi kliknutím na hvězdičku ★. Při zobrazení lze zobrazit pouze jeden vybraný jazyk, nebo všechny.



Obrázek 4 Historie

Čtení a poslech překladu

Po kliknutí na ikonu 🗣️ se přehraje vyslovená verze překladu. Jelikož Seznam nepodporuje překlad celých vět, slova nepřerývá stroj, ale nahraný lidský hlas.

¹ Seznam [online]. [cit. 2016-03-20]. Dostupné z: <https://www.seznam.cz/>

Shrnutí

Popsané slovníky jsou již v provozu delší dobu, jsou tedy schopny více věcí, než zprostředkovat „pouhý“ překlad. Provedl jsem analýzu toho, jak profesionální slovníky fungují a pracují. Těmito slovníky jsem se inspiroval, ať už po funkční nebo grafické stránce a poznatky uplatnil. Stejně jako překladač Seznam [10], můj slovník nepodporuje překládání celých vět, protože tato funkce vyžaduje znalost větné stavby. Využití poslechu je možné uplatnit, protože jej podporuje formát slovníku, avšak velikost slovníku by narostla do ještě větších rozměrů. Překladač Google [9] i můj slovník podporují možnost přidání překladů.

3. Crowdsourcing

Crowdsourcing lze přeložit jako komunitní spolupráce. Je to termín, který označuje využití komunity osob. Například k vyřešení problému, nebo pro tvorbu jakéhokoli obsahu. Nejznámějším příkladem je wikipedie.

Výhody:

- Cíle je možné dosáhnout s poměrně nízkými náklady.
- Je možné zapojit větší rozsah osob, který si sama může firma dovolit zaměstnat.

Šance na selhání:

- Malý počet zúčastněných.
- Nižší kvalita práce.
- Nedostatek motivace pro projekt.
- Jazyková bariéra.

Hlavní překážkou je přesvědčit komunitu uživatelů, aby nám dodávala svá data. Je tedy potřeba, aby uživatelé měli sdílení co nejsnadnější. Dobrý způsob je vytvořit prostředí, které by pro ně bylo zábavné.

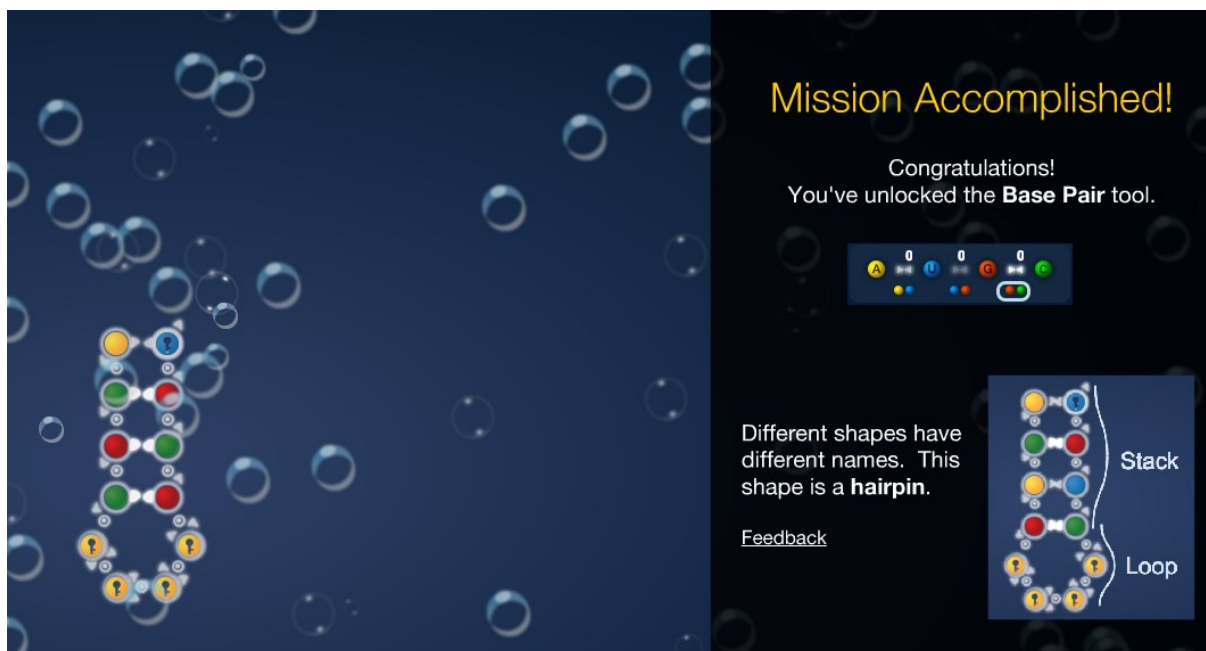
3.1. Hry s účelem

Neboli v angličtině games with purposes, někdy také serious games. Cílem těchto her není pouze zábava, ale možnost se naučit něco nového, popřípadě vyřešit nějaký problém. Hra se učí, tedy čím déle se hraje, tím více mění své chování. Když více uživatelů zareagují jistým způsobem, tak se předpokládá, že je to správný způsob k vyřešení problému. Pro příklad jsou zde uvedené dvě hry.

3.1.1. EteRNA

Cílem projektu EteRNA¹ [4] je, že lidé, kteří nejsou biologové, popřípadě nejsou zvyklí řešit vědecké problémy (nemají stereotypy), mohou vidět vzory, které uniknou odborníkům (Obrázek 5). Konečným cílem je vytvořit první knihovnu rozsáhlých syntetických vzorů RNA (ribonukleová kyselina), které mohou pomoci k vymýcení nemocí. V této online hře hráč navrhuje RNA sekvence, základní stavební bloky života. Každý týden jsou vyhlašování vítězové soutěže. Jsou ohodnoceni podle toho, jak jsou jejich modely složeny a nakonfigurovány podle pravidel syntézy a to dle predikcí počítačového modelu.

¹ Eterna Game [online]. [cit. 2016-03-18]. Dostupné z: <http://www.eternagame.org/web/>



Obrázek 5 příklad úkolu Eterna

3.1.2. Astro Drone¹ [5]

Projekt Astro Drone vytvořila Evropská kosmická agentura (ESA). Tato hra pomůže zjistit, jestli je možné zjistit vzdálenost objektů, pouze pomocí rozdílných pohledů na ně. Jinými slovy: dokáže robot zjistit vzdálenost od objektu, bez použití laseru, sonaru a podobných technologií? Klade si také za cíl zlepšit autonomní schopnosti kosmických sond, jako je přistávání, vyhýbání se překážkám a dokování.



Obrázek 6 Dokování na ISS

¹ European space agency [online]. [cit. 2016-03-18]. Dostupné z: <http://www.esa.int/gsp/ACT/ai/projects/astrodrone.html>

Pro účastnění v tomto projektu je potřeba vlastnit kvadrokoptéru Parrot AR.Drone. Hráči jsou vyzváni, aby provedli různé vesmírné mise v rozšířené realitě (augmented reality).

3.2. Made

Made je jeden z úspěšných crowdsourcingových obchodních modelů, který vymyslel pan Ning Li¹[6]. Vytvořil možnost návrhu vlastního nábytku, výroby a dodání. Pokud budou navržený nábytek zákazníci kupovat, autor dostane 5% za licenční poplatky. Je možné navrhnout židle, stoly, postele, světla a podobně. Firma byla založena v roce 2010 ve Velké Británii a podařilo se jim pokrýt počáteční náklady během jednoho roku.

3.3. Duolingo

Duolingo² [7] je bezplatná internetová stránka pro výuku jazyků. Autorem projektu Duolingo je Luis von Ahn, který je známý jako tvůrce reCAPTCHA³ [8] (viz kapitola 3.4.). Projekt Duolingo funguje od června 2012.

Uživatel si vybere jazyk, který ovládá a podle toho se zobrazí podporované jazyky. Nabízí velké množství jazyků, od španělštiny po waleštinu a další jazyky jsou postupně přidávány. Mezi exotickými jazyky také nabízí esperanto a klingonštinu. Student prochází několika úrovněmi daného jazyka, přičemž jsou mu předkládány překlady, u kterých se předpokládá, že je zvládne. Pokud si však student není jist, že danou větu přeložil korektně, může si překlady konkrétních slov vyhledat ukázkou kurzoru myši na slovo. Systém také umožňuje uživateli hodnotit překlady jiných uživatelů. Tímto způsobem se student může dále poučit o tom, jaké jsou jiné možnosti překladu.

Pro porovnání, jak moc se toho uživatel naučil, slouží mince. Čím více se toho uživatel naučí, tím více mincí má. Tyto mince získá splněním úkolů a testů. Mince slouží k porovnání, jak si uživatel vede vzhledem k ostatním. Každý jazyk má mince zvlášť.

Další službou je překládání textů (Obrázek 7). Překlady fungují podobně jako *reCAPTCHA*, tedy jestliže se více lidí shodne na stejném překladu jedné věty v článku, je větší pravděpodobnost, že překlad je správný (a kvalitnější než strojový překlad).

¹ *Made* [online]. [cit. 2016-04-06]. Dostupné z: <http://www.made.com/about-us>

² *Duolingo* [online]. [cit. 2016-03-10]. Dostupné z: <https://www.duolingo.com/>

³ What is reCAPTCHA. *ReCaptcha* [online]. [cit. 2015-11-10]. Dostupné z: <https://www.google.com/recaptcha/intro/index.html>

Rate or edit this translation

Gigliola Cinquetti

Gigliola Cinquetti (Verona, 20 dicembre 1947) è una cantante, attrice e conduttrice televisiva italiana.

Gigliola Cinquetti

Gigliola Cinquetti nel 1966

Gigliola Cinquetti nel 1966

Nazionalità ItaliaItalia
(Verona)

← BACK

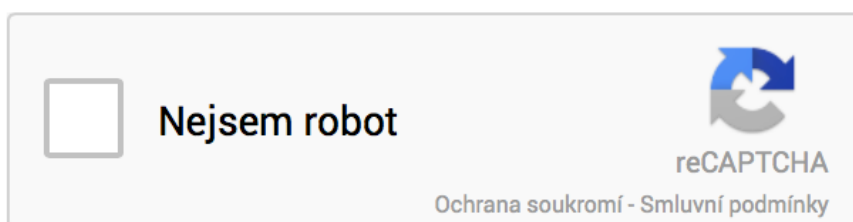
Gigliola Cinquetti (Verona, December 20, 1947) is an Italian singer, actress and TV presenter.

Report abuse Last edit by: Fantatico 207

Obrázek 7 Kontrola překladu textu

3.4. reCaptcha

ReCaptcha [8] poskytuje službu, jejíž hlavní funkcí je blokovat automatické registrace. Například, když je potřeba zajistit, že se nesnaží k určité funkci dostat robot (Obrázek 8). Jinými slovy zajišťuje, že u zařízení je člověk. Jeho používání pomáhá zapojením komunity uživatelů k digitalizaci textů, které počítače nemohou rozpoznat.



Obrázek 8 ReCaptcha kontrola

V případě, že jsou uživatelé přihlášení k Google účtu, jsou automaticky identifikováni. Pokud zde nastane problém, tak budou zobrazeny uživatelům zkomolené znaky, které musí opsat (Obrázek 9) [2]. Tato funkce využívá dvě skupiny znaků. Znak, pro který systém zná správný „překlad“ a znak, pro který správný „překlad“ nezná. Uživatel je požádán, aby napsal obě tyto slova. Pokud napíše slovo, které je již známo správně, systém bude předpokládat, že je správně i druhé slovo. Systém toto opakuje u více uživatelů, aby se zvýšila pravděpodobnost správnosti odpovědi.



Obrázek 9 ReCaptcha zkomolené znaky

Shrnutí

Velmi inspirativní pro mě bylo zjištění, jak velký potenciál v sobě má komunita osob, která spolupracuje na překladu knih, nebo na jejich převedení do digitálního formátu. Informace, že i mezinárodní instituce, jako je Evropská kosmická agentura využívá komunitu k zdokonalení vesmírných sond je překvapivá. Doufám, že i můj slovník také najde uplatnění a bude se nadále rozvíjet.

4. XDXF

XDXF je zkratka pro XML Dictionary Exchange Format¹ [11], a jak jeho název napovídá, určuje formát dat slovníku. Některé další slovníky podporují XDXF, například GoldenDict²[12] a XDClient³[13]. Tento formát slovníku umožňuje klíčovým slovům přiřadit definice. K více klíčovým slovům může být přiřazena jedna či více definic. I když tyto definice mohou obsahovat sebe sama, tuto možnost v práci nevyužívám.

Klíčové slovo může obsahovat:

- `def` počet lidí kteří hlasovali (definition).
- `freq` zpracované hodnocení překladu (frequency).
- `user_freq` uživatelské hodnocení překladu (user frequency).
- `invalidate` pro překlad, který uživatele určili jako špatný.
- `del` kolik uživatelů si přeje smazat překlad (delete).
- `kref` jednoduchý odkaz na jinou klíčovou frázi.
- `gr` gramatické informace o slově (grammatical Information).
- `tr` označuje přepis / výslovnost (transcription/pronunciation).
- `dtrn` přímé překlady (direct translation).
- `etm` etymologické informace o slově (etymological information).
- `categ` text kategorie překladu (category).
- Seznamy:
 - `iref` odkazy na zdroj z internetu (internet reference).
 - `ex` označuje texty příkladů (example).
 - `sr` sémantické vztahy k jiným slovům (semantic relations).

Formát XDXF má mnohem více možností než momentálně používám. Avšak pro efektivní funkčnost jsem se omezil pouze na výše popsané. Podrobnější popis je možné nalézt v Příloha B. Aby bylo možné zajistit hodnocení překladů, bylo potřeba přidat několik dalších atributů. Část z těchto atributů již byly popsány v předchozí práci [2], ale pro úplnost je uvedu i zde.

4.1. Úpravy pro uživatelskou rozšiřitelnost

XDXF popisuje několik doporučení, které by se měly dodržet. Každý slovník je umístěn ve své vlastní složce, a kde název složky je použit jako ID. Musí obsahovat pouze znaky latinské abecedy a nesmí obsahovat mezery nebo jiné speciální znaky. Soubor obsahující samotný slovník musí být pojmenován "`dict.xdx`". Já však podle těchto zásad nepostupuji. Slovník, který uživatelé mohou nahrát na

¹ XDXF Description. *Github* [online]. [cit. 2016-02-01]. Dostupné z: https://github.com/soshial/xdxf_makedict/blob/master/format_standard/xdxf_description.md

² GoldenDict [online]. [cit. 2016-02-02]. Dostupné z: <http://goldendict.org/>

³ XDClient [online]. [cit. 2016-02-02]. Dostupné z: <http://savannah.nongnu.org/projects/xdclient/>

server, může být pojmenován a lokálně uložen libovolně. Aby uživatelé mohli reagovat na přesnost překladů, tak bylo potřeba přidat definicím více atributů.

Jelikož tento typ formátování XML byl definován v DTD, tak bylo potřeba jej prvně převést do XML schématu¹ [14]. Tento převod nebyl kompletní, protože XSD mnohem přesněji popisuje strukturu XML dokumentu.

```
<!ELEMENT def
((gr?,tr*,rref*,def+,ex*,sr?,etm?,categ*)|(gr?,tr*,(#PCDATA,dtrn|kref|rref|iref|abbr|c|ex|co|i|b|gr)*,sr?,etm?,categ*))>
```

XML 1 Úprava v DTD

V ukázce (XML 1) bylo potřeba odstranit element `<sr>`, protože překladač nevěděl, který element má použít, je tedy nejednoznačný (ambiguous).

contributors

Atribut `contributors` (XML 2) značí počet uživatelů, kteří hlasovali konkrétní překlad. Čím více hlasů, tím je větší šance pro to, že je překlad doopravdy přesný. Dle logiky, každý překlad musí mít alespoň jeden hlas.

```
<xs:attribute name="contributors">
  <xs:simpleType>
    <xs:restriction base="xs:unsignedShort">
      <xs:minInclusive value="0"/>
    </xs:restriction>
  </xs:simpleType>
</xs:attribute>
```

XML 2 Atribut `user_freq` v XSD

freq

`freq` atribut je již implementován v originálním standardu, s významem absolutní nebo relativní četnosti hodnoty definice. Zde je použit jako vypočtený aritmetický průměr všech uživatelských hlasování o přesnosti překladu (XML 3). Atribut může nabývat hodnot od 0 do 10.

- 0: překlad je velice nepřesný.
- 10: překlad odpovídá nejvíce realitě.

```
<xs:attribute name="freq">
  <xs:simpleType>
    <xs:restriction base="xs:double">
      <xs:minInclusive value="0"/>
      <xs:maxInclusive value="10"/>
    </xs:restriction>
  </xs:simpleType>
</xs:attribute>
```

XML 3 Atribut `user_freq` v XSD

¹ MLÝNKOVÁ, Irena, POKORNÝ, Jaroslav. *XML technologie: principy a aplikace v praxi*. 1. vyd. Praha: Grada, 2008, 267 s. Průvodce (Grada). ISBN 978-80-247-2725-7

user_freq

Atribut `user_freq` určuje uživatelské hodnocení přesnosti překladu (XML 4). Může nabývat hodnot od -1 do 10.

- -1: překlad je nesprávný a chci, aby byl odstraněn z oficiálního slovníku. Použije se tehdy, když se v oficiálním slovníku nalezne překlad, který považován za špatný.
- 0: překlad je velice nepřesný.
- 10: překlad odpovídá nejvíce realitě.

```
<xs:attribute name="user_freq" >
  <xs:simpleType >
    <xs:restriction base="xs:int">
      <xs:minInclusive value="-1"/>
      <xs:maxInclusive value="10"/>
    </xs:restriction>
  </xs:simpleType>
</xs:attribute>
```

XML 4 Atribut user_freq v XSD**del**

Tento atribut `del` (XML 5) poukazuje na to, kolik uživatelů hlasovalo pro smazání překladu. Pokud číslo překročí mez určenou správcem pro daný slovník, bude překlad smazán. Atribut je pouze pro interní použití, protože uživatel nemá nutnost vědět počet hlasujících uživatelů pro smazání překladu.

```
<xs:attribute name="del">
  <xs:simpleType >
    <xs:restriction base="xs:unsignedShort">
      <xs:minInclusive value="0"/>
    </xs:restriction>
  </xs:simpleType>
</xs:attribute>
```

XML 5 Atribut del v XSD**Invalidate**

Atribut `invalidate` (XML 6) má význam pouze pro správce, je tedy stejně jako `del` interní. Dle tohoto atributu se rozhoduje, jestli bude překlad zapsán do souboru se slovníkem, který je určen pro uživatele ke stažení.

- True: Překlad je špatný, při tvorbě uživatelského slovníku bude ignorován.
- False: Překlad je v pořádku. V tomto případě se atribut nezapisuje.

```
<xs:attribute name="invalidate">
  <xs:simpleType >
    <xs:restriction base="xs:boolean">
    </xs:restriction>
  </xs:simpleType>
</xs:attribute>
```

XML 6 Atribut invalidate v XSD

Shrnutí

Používání formátu slovníku XDXF je výhodné, poněvadž je již k dispozici delší dobu; poslední aktualizace formátu proběhla v roce 2015¹ [15]. I v případě, že slovník používá pouze část XDXF formátu, lze použít i slovníková data z ostatních programů podporujících tento formát. Opačným způsobem to nelze, kvůli přidání nových atributů k definicím pro podporu jejich hodnocení, `user_freq` a další. Formát slovníku se liší mezi slovníkem, který je určen pro uživatele a slovníkem, který je pouze interní, tedy pro správce. V programu nejsou využívány všechny možnosti, které XDXF formát nabízí. Ty by bylo možno dodat po žádostech uživatelů, v případě, že jsou obeznámeni s možnostmi tohoto formátu.

¹ *Source Forge* [online]. [cit. 2016-03-10]. Dostupné z: <https://sourceforge.net/projects/xdxf/>

5. Architektura & návrh

Minimální systémové požadavky serveru vzhledem k operačnímu systému

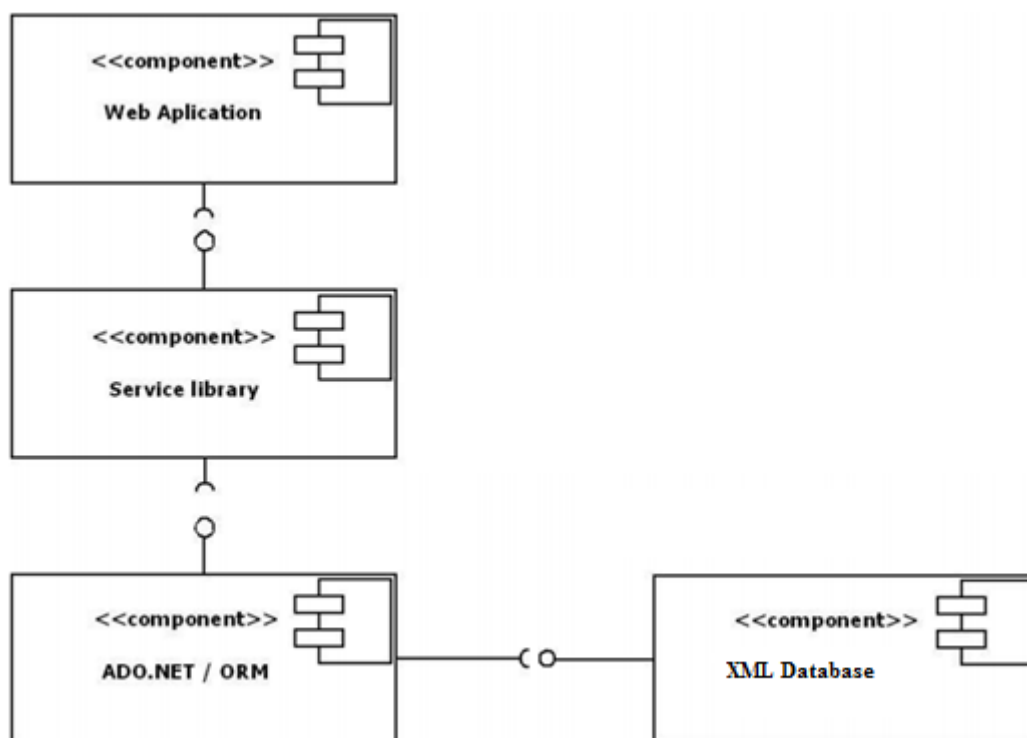
- Procesor 1,4 GHz 64 bitový¹ [16]
- Volné místo na disku 4,5 GB (doporučuje se více: 8,5 GB pro všechny konfigurace).
- 2 GB fyzické paměti RAM (doporučuje se 4 GB).
- Jelikož se předpokládá, že se načte do paměti celý slovník, tak se nároky na paměť RAM mohou zvýšit. Vzhledem k tomu, že při načtení slovníku s přibližným počtem klíčových slov 87 400 je paměť zatížena alespoň 1GB, tak by bylo dobré, aby server měl alespoň 4 GB.

Webová aplikace:

Diagram komponent, zobrazující komponenty serveru a závislosti mezi nimi (Obrázek 10). Webová aplikace přistupuje do databáze se pomocí obslužné knihovny, tedy ne napřímo.

ADO.NET/ vlastní ORM

ASP.NET / C#

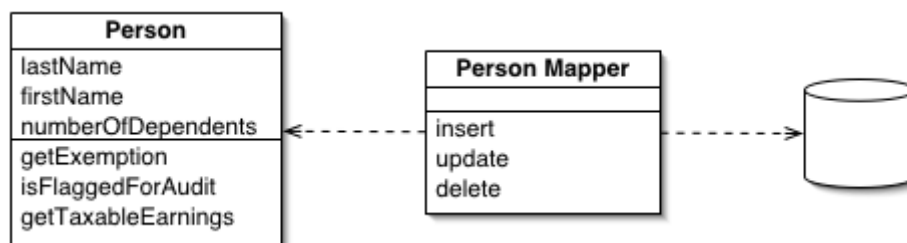


Obrázek 10 Diagram komponent serveru

¹ .NET Framework System Requirements [online]. [cit. 2016-03-22]. Dostupné z: <https://msdn.microsoft.com/en-us/library/8z6watww%28v=vs.100%29.aspx>

Přístup k online slovníku je realizován pomocí návrhového vzoru data mapper¹ [17] (Obrázek 11). Vrstva mapperů, které přesouvají data mezi DB a objekty přičemž jsou k sobě navzájem a mapperům nezávislé. Tedy je to vrstva, která odděluje objekty v paměti od databáze.

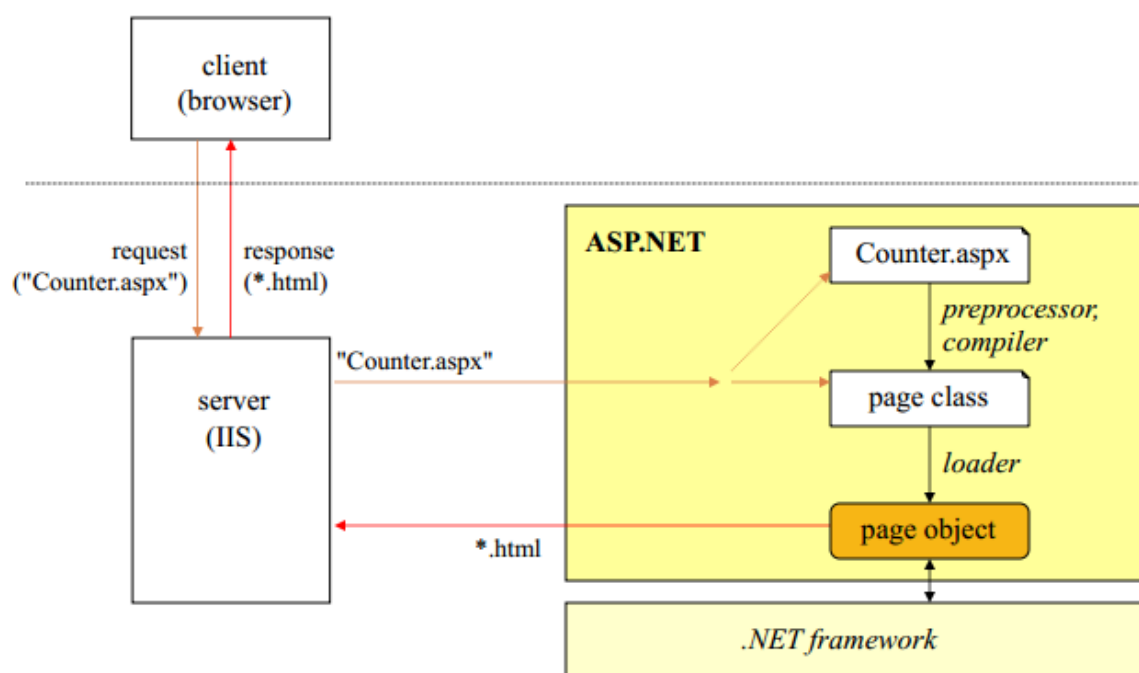
Výhoda: Dobře rozdělené kompetence, vzájemná izolace objektů v paměti.



Obrázek 11 Data Mapper

5.1. ASP.NET

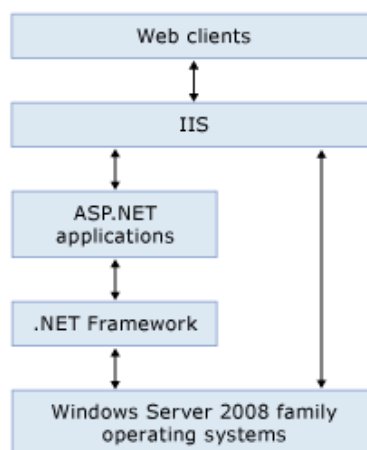
Stránky jsou založeny na ASP.NET. Obrázek 12 ukazuje interní funkcionalitu (behind the scenes). Klient pošle požadavek na server, který jej zpracuje a pošle zpět odpověď klientovi.



Obrázek 12 Funkčnost asp.net²[18]

¹ FOWLER, Martin. *P of EAA* [online]. 2000 [cit. 2016-02-25]. Dostupné z: <http://martinfowler.com/eaCatalog/dataMapper.html>

² MARTINOVIČ, Jan. *Overview .Net*. Dostupné také z: <http://ita.vsb.cz/11p%5%99edn%C3%A1%C5%A1ky.aspx>



Obrázek 13 Architektura asp.net

Jak ukazují obrázky (Obrázek 12, Obrázek 13), všichni webovní klienti (tzn. webový prohlížeč použitý uživatelem slovníku) komunikují s aplikacemi ASP.NET pomocí Microsoft Internet Information Services (IIS)¹ [19]. V případě potřeby IIS požadavek ověří a poté vyhledá požadovaný zdroj (například aplikace ASP.NET). Je-li klient oprávněn, je pro něj zdroj k dispozici.

5.1.1. Použité prvky

K uložení slovníku na server poslouží třída `FileUpload`²[20], jenž zajistí vše potřebné pro uložení slovníku. Zobrazí okno, přes které uživatel vybere soubor k nahrání.

Přihlašování je vyřešeno pomocí zakomponovaného systému v ASP.NET³[21]. Obsahuje přihlašování, odhlašování, role a jejich oprávnění atd.

5.2. Role

Na serveru jsou definovány tři role (Obrázek 14):

5.2.1. Anonymní uživatel

Kvůli bezpečnosti může tato role pouze stahovat slovníky. Nahrávání slovníků je zakázáno. Nepřihlášený uživatel se může přihlásit, popřípadě si vytvořit nový účet.

¹ Microsoft MSDN: *ASP.Net security Architecture* [online]. [cit. 2016-03-12]. <https://msdn.microsoft.com/en-us/library/yedba920.aspx>

² Microsoft MSDN: *FileUpload Class* [online]. [cit. 2016-03-12]. <https://msdn.microsoft.com/en-us/library/system.web.ui.webcontrols.fileupload%28v=vs.110%29.aspx>

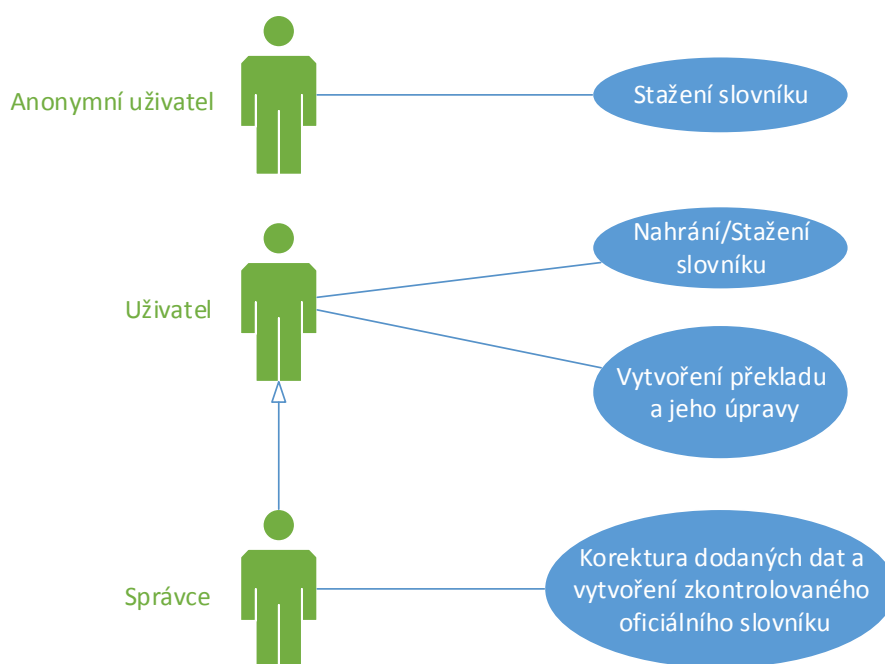
³ Microsoft MSDN : *Administration* [online]. [cit. 2016-04-10]. Dostupné z: <https://msdn.microsoft.com/en-us/library/ssa0wsyf.aspx>

5.2.2. Přihlášený uživatel

Uživatelé po přihlášení mohou nahrát svůj uživatelský slovník na server. Pomocí nahraného slovníku bude vytvořen nový nezkontrolovaný slovník, který si posléze může stáhnout.

5.2.3. Správce

Hlavní funkcí správce je kontrolovat překlady a tím vytvářet přesnější slovník, viz kapitola 5.3.



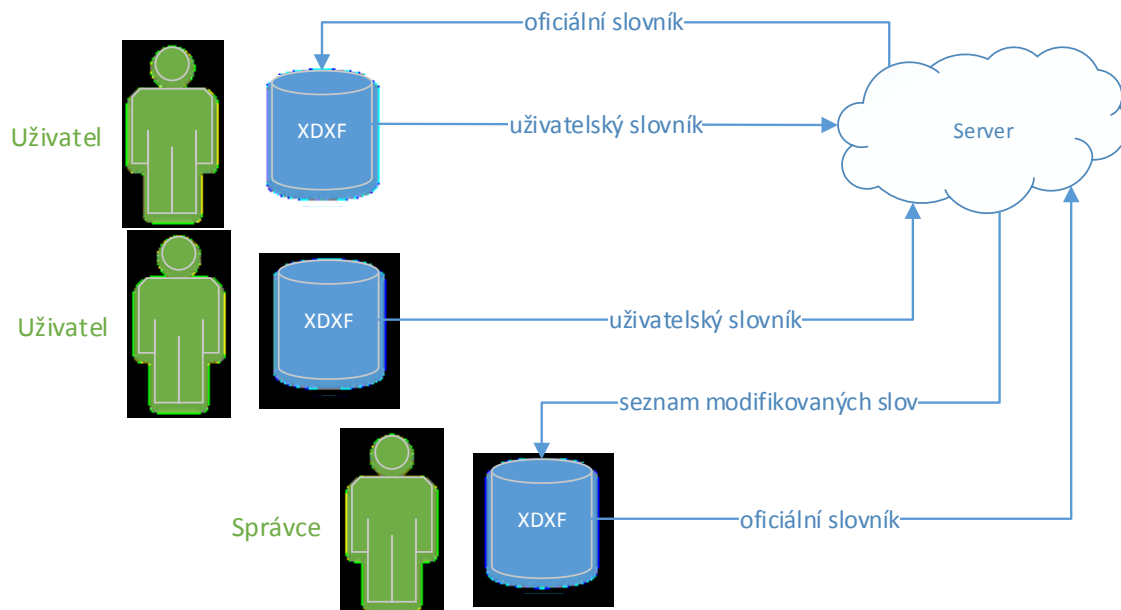
Obrázek 14 Příklad užití, práva

5.3. Kontrola kvality

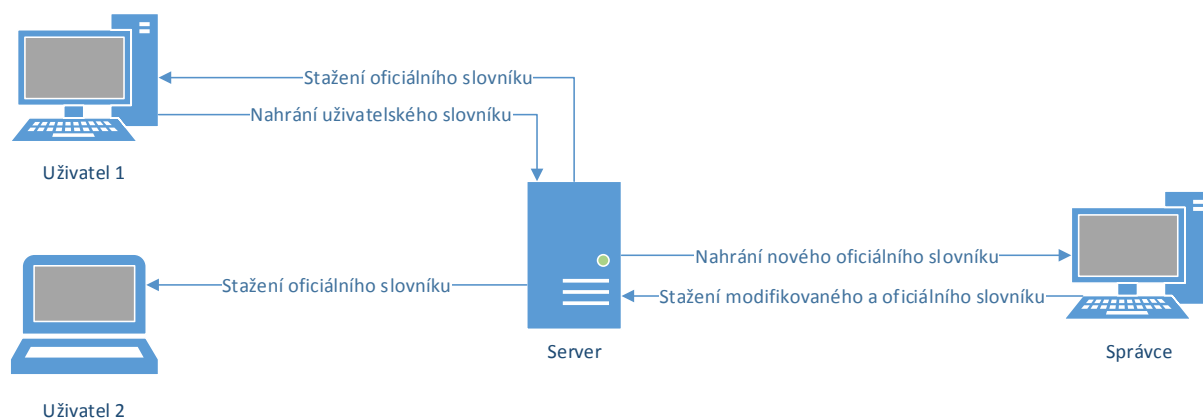
U strojového spojení slovníků (oficiální s uživatelským) hrozí výskyt chyb. Strojové spojení znamená spojení dvou slovníků bez ohledu na správnost dat. V zodpovědnosti role správce je provádění korektur (Obrázek 14). Správce bude pravidelně procházet pouze upravená/nová slova a kontrolovat je. Po takové kontrole správce aktualizuje verzi oficiálního slovníku a poskytne ho uživatelům ke stažení (Obrázek 15, Obrázek 16).

Na serveru bude ke stažení k dispozici:

- Strojově spojená verze slovníku nezkontrolovaná pověřenou osobou; možné nesrovnalosti.
- Ověřený oficiální slovník.
- Komprimovaná podoba obou slovníků.



Obrázek 15 Aktualizace slovníku



Obrázek 16 Komunikace se serverem při sdílení slovníku

5.4. Formální vytipování místa na trhu (cílová skupina)

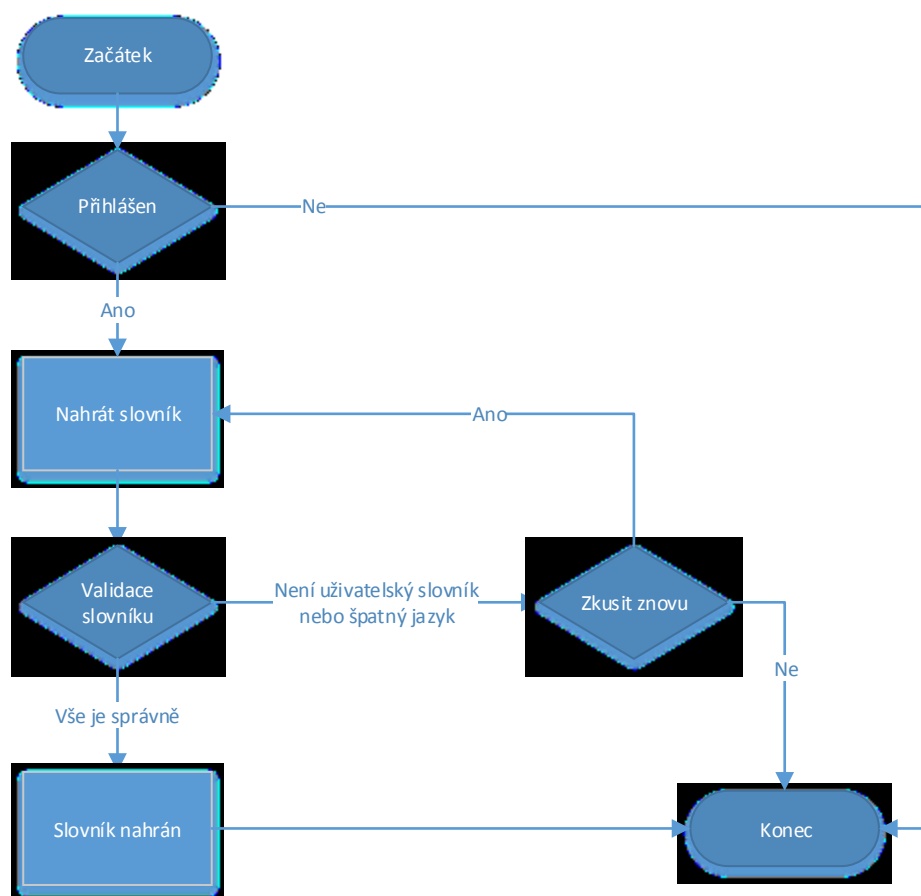
Slovník je určen převážně pro uživatele, kteří budou ochotni sdílet svůj slovník pro jeho rozvoj. Po patřičném zapojení lidí v určitém oboru lze také vytvořit lékařský, stavební a jiné specializované slovníky z těchto oborů. Slovník je určen pouze pro jednotlivá slovíčka a fráze, ne pro celé věty.

5.5. Životní cyklus

Slovník a jeho data mohou být modifikována. Níže jsou popsány hlavní modifikace.

5.5.1. Nahrání slovníku

Pro nahrání slovníku musí být uživatel přihlášen a hodnoty slovníku `lang_from` a `lang_to` musí odpovídat nahrávané stránce. Princip algoritmu ilustruje Obrázek 17. Nelze tedy nahrát Anglicko-Německý slovník do Anglicko-Českého. Také je možné spojit pouze uživatelské slovníky. Tedy ten, který má atribut `publisher` v `meta_info` nastaven na `User`.



Obrázek 17 Nahrání slovníku na server

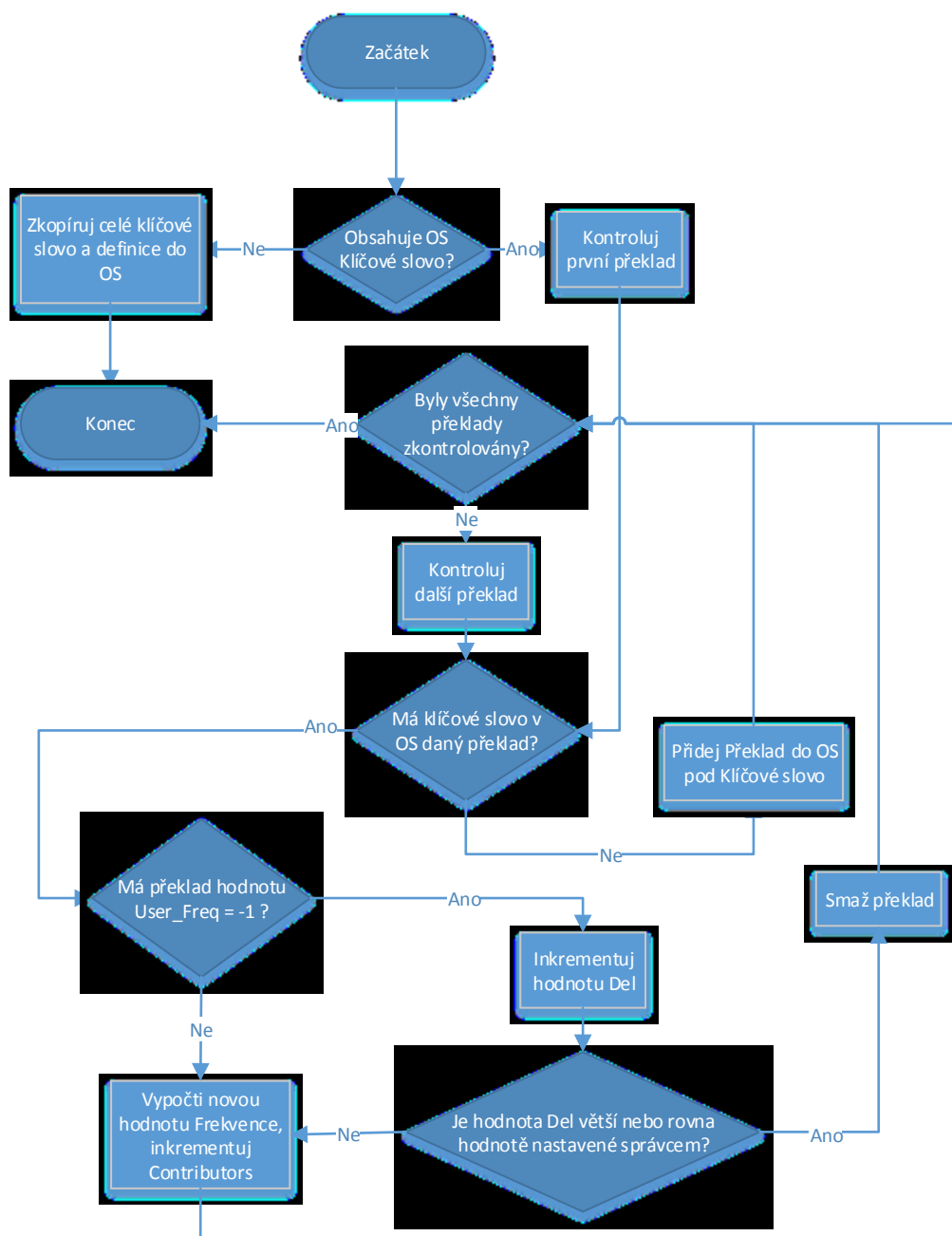
5.5.2. Spojení slovníků

Nahráný slovník (Obrázek 18, zkratka OS značí oficiální slovník) se sloučí se slovníkem oficiálním. Při spojování slovníků se porovná, zda nezkontrolovaný oficiální slovník již neobsahuje daný překlad. Pokud jej neobsahuje, je do něj zkopírována celá definice. V případě, že oficiální slovník daný překlad obsahuje, tak se kontrolují definice. Jestli oficiální slovník obsahuje danou definici a tato definice je uživatelsky ohodnocena, tak je aktualizována hodnota `Freq` dle Rovnice 1:

$$freq_{new} = \frac{(freq_{old} * contributors) + user_freq}{(contributors + 1)}$$

Rovnice 1 Nová frekvence po automatickém spojení

V případě, že definice není známa, tak je zkopírována pod dané klíčové slovo. Zároveň se provádí kontrola, kolik uživatelů chce smazat daný překlad, když je jich dostatek (určeno správcem), tak je překlad smazán. V nezkontrolovaném oficiálním slovníku je rovněž inkrementována hodnota *contributors*, která označuje počet uživatelů, kteří hodnotili překlad.



Obrázek 18 Spojení překladů

Poté se kontroluje, jestli již není daná definice v dočasném slovníku, jež obsahuje pouze data poskytnutá uživateli. Liší se však v tom, že se zde kontrolují i ostatní atributy překladu, jako například gramatické informace a výslovnost. V případě, že jsou identické, postupuje se stejně, jak je popsáno výše. Když se liší, tak je vytvořena nová definice, kam se tyto data zkopírují.

5.5.3. Ověření správcem

Při kontrole správce se prochází slovník, který obsahuje pouze modifikované data. Jinými slovy tedy slovník, vytvořený pouze z uživatelských slovníků. Správce se rozhoduje, jestli daný překlad, popis, hodnocení atd. použije. Jestli ano, tak nová $Freq$ se vypočte daným způsobem, který popisuje Rovnice 2:

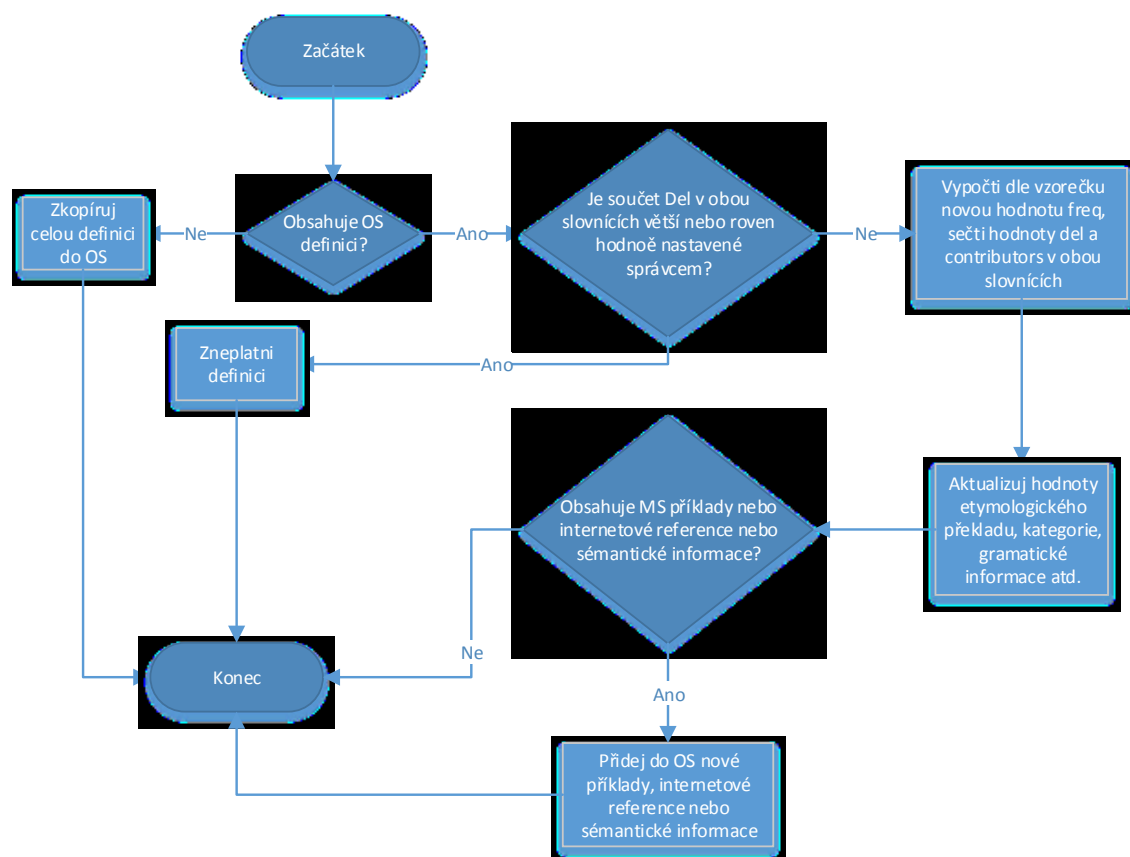
$$Freq_{new} = \frac{(freq_{ofi} * contributors_{ofi}) + (freq_{mod} * contributors_{mod})}{contributors_{ofi} + contributors_{mod}}$$

Rovnice 2 Výpočet nové frekvence po potvrzení správcem

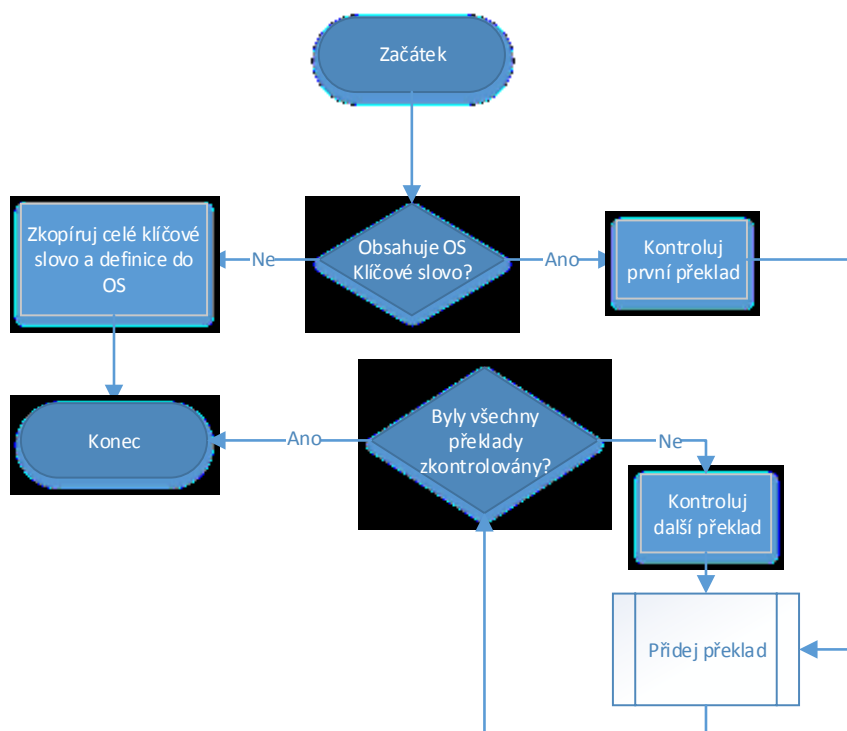
Pokud má daný překlad součet atributu `del` větší než hodnota v konfiguračním souboru, tak je tento překlad smazán. V případě, že klíčové slovo nebude mít alespoň jeden překlad, je dané klíčové slovo smazáno ze slovníku. Toto ovšem platí pouze pro slovník, který je přístupný uživatelům. Interní slovník bude obsahovat i překlady a klíčová slova, která byly smazány ze slovníku. Tyto překlady budou pouze zneplatněné označením atributem `invalidate`.

Při procházení modifikovaného slovníku má správce na výběr:

- **Vypočti nové hodnoty $Freq$, $Contributors$ a Del :** Správce přidá hodnoty daného překladu, do oficiálního slovníku (dle předchozí rovnice nahoře), sečte se hodnota $Contributors$ a Del .
- **Přidej překlad (Obrázek 19):** Jsou vypočteny nové hodnoty $Freq$, $Contributors$ a Del a případně zneplatní překlad. Přidají se nové internetové reference, příklady a sémantické informace. Aktualizují se hodnoty etymologického překladu, kategorie, gramatické informace atd.
- **přidat celé klíčové slovo (Obrázek 20):** Tímto se projde oficiální slovník a porovná se, jestli existuje klíčové slovo i v modifikovaném slovníku. Když ne, tak se přidá celé do slovníku. Když ano, tak se porovnávají definice a postupuje se stejně, jak je popsáno dříve.



Obrázek 19 Přidej překlad



Obrázek 20 Přidej klíčové slovo

5.6. Komprimace souborů

Slovník o 87 000 klíčových slovech má velikost začínající na 10 MB. Tato velikost bude narůstat s počtem překladů a atributů. Proto by bylo výhodné mít komprimovanou verzi slovníku ke stažení, přestože uživatel bude muset soubor se slovníkem „rozbalit“. Tato možnost je pouze doplňková, protože rychlost stahování dat je dostatečně velká pro rychlé stažení souboru.

Pro komprimaci je výhodné použít formát gzip¹ [22] protože jej podporuje VisualStudio, takže není nutnost instalovat komprimační moduly, popřípadě kupovat licenci. Díky tomu, že slovník je ve formátu XML, můžeme na něj pohlížet jako na textový soubor, který obecně jde efektivně komprimovat. Po komprimaci tento soubor má velikost přibližně desetkrát menší, tedy z 15MB na 1,5MB.

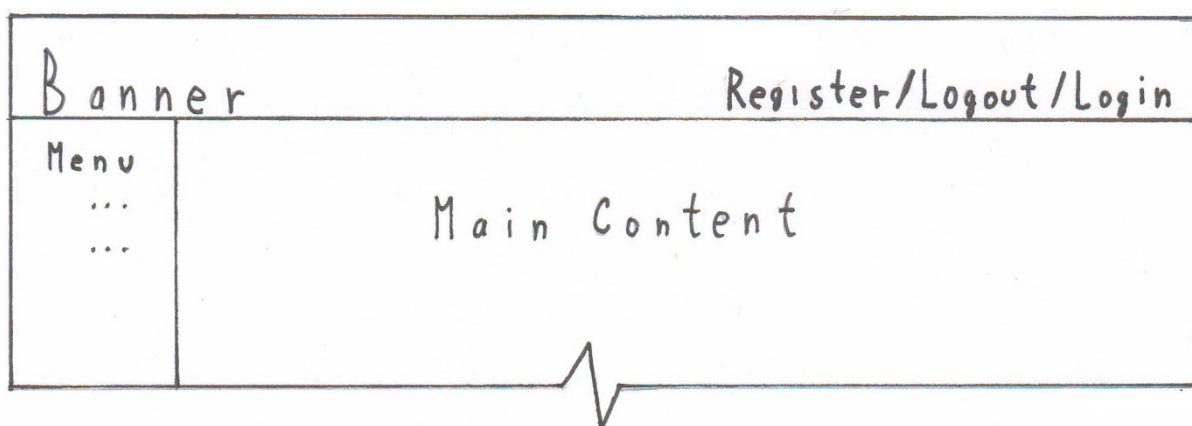
¹ GZip [online]. [cit. 2016-03-10]. Dostupné z: <http://www.gzip.org/>

6. Uživatelské rozhraní stránek

Jelikož tyto stránky slouží primárně jako prostředník, přes kterého si uživatelé stáhnou slovník, uživatelské rozhraní je navrženo a vytvořeno co nejjednodušší. Pro uživatele nejdůležitější funkcionality je stažení oficiálního slovníku a nahrání uživatelského slovníku.

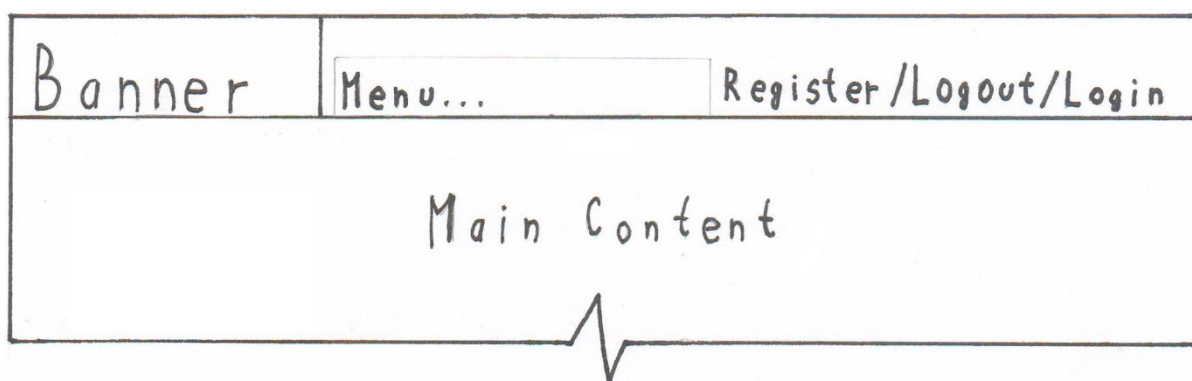
6.1. Master Page

Master page je jednoduché GUI pro hlavní stránku, do které se budou vkládat jednotlivé stránky. Připravil jsem dvě varianty. Rozdíl je v umístění menu, jedno horizontální na začátku stránky (Obrázek 22) a druhé vertikální po levé straně (Obrázek 21).



Obrázek 21 Vertikální menu

Z těchto variant je naimplementováno Master Page s horizontálním menu, protože je potřeba více místa vzhledem k šířce stránky (kapitola 6.3).



Obrázek 22 Horizontální menu

Dle potřeby by bylo možné Master Page změnit za druhou variantu, popřípadě by se mohlo vytvořit úplně jiné rozmístění prvků.

6.2. Stránky slovníků

Tato stránka je záměrně co nejjednodušší (Obrázek 23). Uživateli se nabídne základní popis slovníku a možnost stažení nejnovějších slovníků daného jazyka. Je na výběr nezkontrolovaný slovník, což je pouze strojové spojení slovníků a správcem zkontrolovaný slovník. Vzhledem k relativně velké velikosti souboru slovníku je možnost stáhnout obě varianty slovník v komprimované formě ve formátu .gz.

Dictionary information: ...

Upload form

Browse

Upload

Last update on: ...

Download Unchecked Official Dictionary

————||————(Compressed)

Download Official Dictionary

————||————(Compressed)

Obrázek 23 Stránka slovníků

6.3. Stránka správce (Administrátora)

Tyto stránky jsou cílené na to, aby správce, nebo někdo kdo bude kontrolovat přesnost překladů, mohl mít možnost kontrolovat přesnost překladů a jejich různých atributů. Stránka je rozdělená na dvě části, v každé jsou zobrazena data z jiných slovníků (Obrázek 24). V levém sloupci jsou data z modifikovaného slovníku, tedy ze spojených uživatelských slovníků. V pravé části jsou data z oficiálního slovníku. Ty jsou zobrazeny, aby správce měl možnost porovnat již existující data překladů a rozhodnout se, jestli přidat celý překlad, nebo pouze jeho část.

Select language <input type="text"/>		<input type="button" value="Load dictionary"/>		<input type="button" value="Save dictionary"/>			
<input type="button" value="Prev"/>		Modified Dictionary		Official Dictionary <input type="button" value="Next"/>			
Key Phrase	<input type="text"/>	<input type="button" value="+"/>		<input type="text"/>			
Definition	<input type="text"/>	<input type="button" value="←"/> <input type="button" value="→"/> <input type="button" value="+"/>		<input type="text"/>	<input type="button" value="←"/> <input type="button" value="→"/>		
Freq	<input type="text"/>			<input type="text"/>			
User freq	<input type="text"/>	<input type="button" value="+"/>		Invalidate True/False			
Contributors	<input type="text"/>			<input type="text"/>			
Delete?	<input type="text"/>			<input type="text"/>			
Direct translation	<input type="text"/>	<input type="button" value="+"/>		<input type="text"/>			
Category	<input type="text"/>	<input type="button" value="+"/>		<input type="text"/>			
Grammar information	<input type="text"/>	<input type="button" value="+"/>		<input type="text"/>			
Ethymologic info.	<input type="text"/>	<input type="button" value="+"/>		<input type="text"/>			
Reference	<input type="text"/>	<input type="button" value="+"/>		<input type="text"/>			
Internet reference	<input type="text"/>	<input type="button" value="←"/> <input type="button" value="→"/> <input type="button" value="-"/> <input type="button" value="+"/>		<input type="text"/>	<input type="button" value="←"/> <input type="button" value="→"/>		
Semantic rel.	kcmt	Text	<input type="button" value="-"/> <input type="button" value="+"/>	Semantic rel.	kcmt	Text	<input type="button" value="←"/>
<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="button" value="←"/> <input type="button" value="→"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="button" value="→"/>
Examples	Type	Text	<input type="button" value="-"/> <input type="button" value="+"/>	Examples	Type	Text	<input type="button" value="←"/>
<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="button" value="←"/> <input type="button" value="→"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="button" value="→"/>

Obrázek 24 Stránka správce

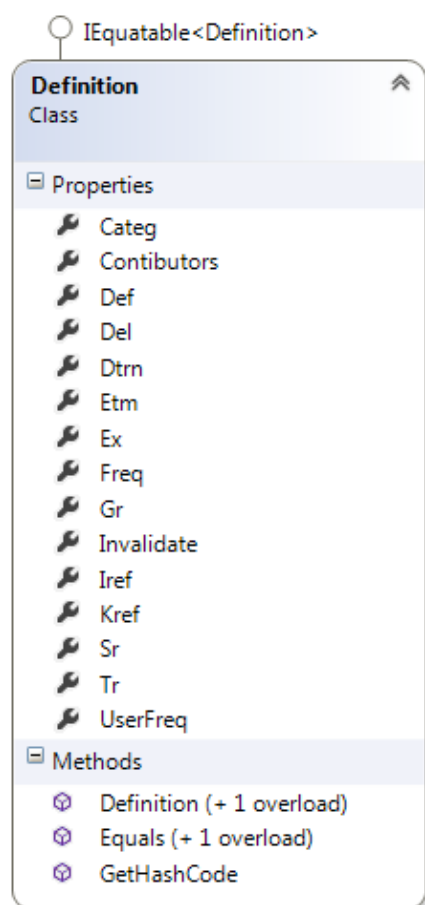
Shrnutí

Grafické rozhraní bylo z hlediska uživatelů navrženo co nejjednodušší. V případě stránky správce je jednoduchost takřka nemožná. Z hlediska rozšíření funkcionality, aby program podporoval více možností formátu XDXF, tak by tato stránka nabyla ještě větší složitější.

7. Implementace

7.1. Překlad

Každé klíčové slovo musí mít alespoň jeden překlad (Obrázek 25). Samotné slovo je spárování klíčového slova s definicí popřípadě definicemi. Ve slovníku je klíčové slovo unikátní, může se vyskytovat pouze jednou. Každý objekt třídy definice představuje určitou část překladu, z toho pouze odkazy na internetové stránky, příklady a sémantické vztahy se mohou vyskytovat vícekrát.



Obrázek 25 Definice

Tato třída pro zjednodušení funkčnosti implementuje možnost využívat porovnávací metodu, proto dědí z rozhraní `IEquatable`.

```
public class Definition : IEquatable<Definition>
```

Zdrojový kód 1 Hlavička Definice

Přepisuje se chování metody `equals`¹ [23] (Zdrojový kód 2), aby byla schopna porovnávat složité objekty třídy `Definition`. Porovnává všechny atributy až na odkazy na číselné ohodnocení překladu (`freq`, `user_freq`, `del`, `contributors`), internetové stránky, příklady a sémantické vztahy. Číselné ohodnocení se neporovnává, protože se tato funkce využívá pouze při jejich spojování. Tedy kdyby se číselné hodnocení porovnávalo, tak by funkce nedokázala rozpoznat stejné překlady.

```

public override bool Equals(object obj)
{
    return this.Equals(obj as Definition);
}
public bool Equals(Definition other)
{
    if (other == null)
        return false;

    return
        (
            this.Def == other.Def ||
            this.Def != null &&
            this.Def.Equals(other.Def)
        ) &&
        . . .
        &&
        (
            this.Dtrn == other.Dtrn ||
            this.Dtrn != null &&
            this.Dtrn.Equals(other.Dtrn)
        )
    }
public override int GetHashCode()
{
    return base.GetHashCode();
}

```

Zdrojový kód 2 Porovnání definice

Protože byla přepsána metoda `equals`, tak je také potřeba přepsat `GetHashCode`. Tuto funkci jsem neupravoval.

7.2. Slovník

Každý slovník obsahuje hlavičku `XDxf`, `MetaInfo` a spárovaný seznam klíčových slov s definicemi (Zdrojový kód 3, Obrázek 26).

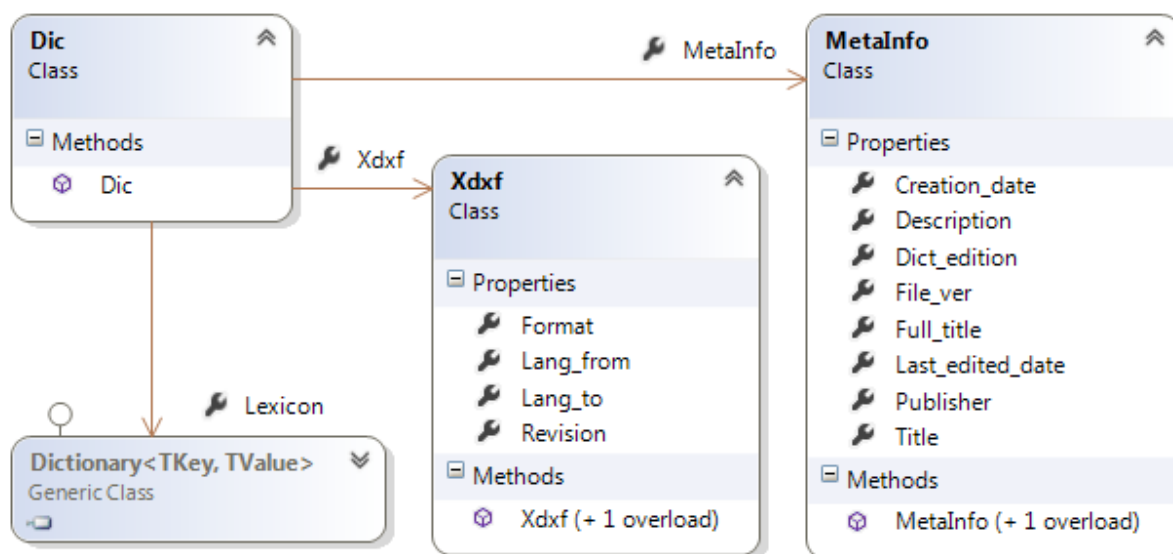
```

public Xdxf Xdxf { get; set; }
public MetaInfo MetaInfo { get; set; }
public Dictionary<string, List<Definition>> Lexicon { get; set; }

```

Zdrojový kód 3 Slovník

¹ *Stack overflow*. [online]. 4.5.2012 [cit. 2016-03-04]. Dostupné z: <http://stackoverflow.com/questions/10454519/best-way-to-compare-two-complex-object>



Obrázek 26 Třída Slovník

Udržují se následující verze slovníků:

- Nezkontrolovaný oficiální slovník a jeho komprimovaná verze.
- Ověřený oficiální slovník a jeho komprimovaná verze.
- Modifikovaný slovník obsahující spojené uživatelské slovníky.
- Oficiální slovník, pouze pro správce.

Slovník, který uživatel bude moci stáhnout, nebude obsahovat atributy `Del` a `Invalidate`. Tyto atributy jsou pouze pro slovník správce.

7.3. Nahrání slovníku a jeho aktualizace

Slovník může nahrát pouze přihlášený uživatel. Každý nahraný slovník je uložen s modifikovaným názvem, ke kterému byla přidána časová známka s přesností na milisekundy, aby se zabránilo nahrání dvou či více slovníků se stejným označením. Při nahrávání slovníku si systém udržuje načtené 3 slovníky (Zdrojový kód 4):

- uživatelsky: nahrává jej uživatel
- modifikovaný: je tvořen spojením uživatelských slovníků
- oficiální

```

App_Code.Dic UserDictionary = App_Code.ControlServices.LoadXML(fromPath);
App_Code.Dic OfiDictionary = App_Code.ControlServices.LoadXML(ofiPath);
App_Code.Dic ModDictionary = App_Code.ControlServices.LoadXML(moded);
  
```

Zdrojový kód 4 načtení slovníků

Postupně se prochází nahraný uživatelský slovník a kontrolují se klíčová slova a definice (Obrázek 18). Do oficiálního slovníku se pouze přidávají a odebírají klíčová slova a definice. Nepřidávají se

přímé překlady, internetové reference a podobně. Do modifikovaného slovníku naopak přidávají veškeré atributy. Oba slovníky vypočítávají novou frekvenci (Rovnice 1) a navyšují počet uživatelů, kteří dané slovo hodnotili.

Protože při použití příkazu `foreach` pro procházení seznamem není možné přidávat, či odebírat prvky tohoto seznamu (obdobné problémy mohou nastat i u funkce `for`), vytváří se kopie seznamu oficiálního slovníku. V této kopii nejsou uloženy pouze odkazy na hodnoty (reference), ale samotné primitivní datové typy. V angličtině se tento typ kopírování nazývá *deep copy*¹. [24] Prochází se tedy překlad v oficiálním slovníku, ale veškeré úpravy se provádí u jeho kopie.

```
foreach (App_Code.Definition to in toRecord)
{
    App_Code.Definition defi = new App_Code.Definition(to);
    modified.Add(defi);
}
List<App_Code.Definition> remove = new List<App_Code.Definition>();
foreach (App_Code.Definition from in fromRecord)
{
    bool copy = true; //překlad neexistuje v oficialnim slovníku
    for (int i = 0; i < toRecord.Count(); i++)
    {
        //existuje definice v ofi slovníku i uživatelském
        if (from.Def == modified[i].Def)
        {
            copy = false;
            modified[i].Freq =
            ((modified[i].Freq * modified[i].Contributors) + from.UserFreq) /
            ++modified[i].Contributors;
            modified[i].Freq = Math.Round(modified[i].Freq, 2);
            //Když je to označeno pro smazání
            if (from.UserFreq == -1)
            {
                modified[i].Del++;
                if (modified[i].Del >= del) //příprava pro smazání
                {
                    remove.Add(modified[i]);
                }
            }
        }
    }
}
.
.
.
```

Zdrojový kód 5 Aktualizace oficiálního slovníku

Do proměnné `copy` se uloží, zda se našla stejná definice v uživatelském i oficiálním slovníku. Když se nenašla, tak se celá definice zkopíruje a uživatelské hodnocení se nastaví jako frekvence. Frekvence se aktualizuje dle vzorce uvedeného v kapitole 5.5.2 (Rovnice 1).

¹ *Net Informations* [online]. [cit. 2016-03-04]. Dostupné z: <http://net-informations.com/faq/net/shallow-deep-copy.htm>

Hodnota `UserFreq=-1` značí, že překlad je špatný a měl by být smazán. Pokud počet hlasů pro smazání překladu bude stejný, nebo překročí hodnotu nastavenou správcem, tak je překlad smazán. Toto ovšem platí pouze pro nezkontrolovaný slovník. Při vytváření modifikovaného slovníku se překlad nemaže, to se stane až tehdy, když se správce rozhodne při kontrole správnosti překladů.

Rozdíl v modifikovaném slovníku je v tom, že se porovnávají i sémantické vztahy, internetové reference a příklady. Jelikož `equals` nedokáže porovnávat seznamy, tak se v každém cyklu zvlášť kontroluje, jestli nejsou seznamy stejné.

```

if (from.Equals(modified[i]))
{
    bool equality = CompareLists.Compare2Lists(from.Iref,
                                                modified[i].Iref);
    bool equality1 = CompareLists.Compare2Lists(from.Sr,
                                                modified[i].Sr);
    bool equality2 = CompareLists.Compare2Lists(from.Ex,
                                                modified[i].Ex);
    if (equality && equality1 && equality2)
    {
        modified[i].Freq = ((modified[i].Freq *
modified[i].Contributors) + from.UserFreq) / ++modified[i].Contributors;
        modified[i].Freq = Math.Round(modified[i].Freq, 2);
        //Kdyz je to oznaceno pro smazani
        if (from.UserFreq == -1) modified[i].Del++;
    }
    else
    {
        copy = true;
    }
}
else
{
    copy = true;
}
}

```

Zdrojový kód 6 Aktualizace modifikovaného slovníku

Při nahrání nového slovníku se automaticky aktualizují 2 slovníky, modifikovaný a oficiální. Tento oficiální slovník je pouze dočasný. Bude nahrazen, až správce vytvoří novou verzi oficiálního slovníku pomocí slovníku modifikovaného.

7.4. Stažení slovníku

Jak již bylo zmíněno (kapitola 6.2), uživatel má možnost stáhnout čtyři verze slovníku (Zdrojový kód 7). Zkontrolovaný, nezkontrolovaný a jejich komprimovaná verze, kterou jsem přidal po doporučení vedoucího práce. Oficiální slovník s 87 000 klíčovými slovy má přibližně 15 MB. Oproti tomu komprimovaná verze má pouze desetkrát méně, tedy 1,5 MB. Při kliknutí na stažení se uživateli buď stáhne slovník do předem vybrané složky, nebo se otevře okno, kde si může vybrat, co se souborem chce udělat. Tedy podle toho co má uživatel nastaveno v prohlížeči.

```

string filePath = Server.MapPath(path);
Response.ContentType = ContentType;
Response.AppendHeader("Content-Disposition", "attachment; filename=" +
Path.GetFileName(filePath));
Response.WriteFile(filePath);
Response.End();

```

Zdrojový kód 7 Stažení slovníku

7.5. Kontrola slovníku správcem

Aby si správce mohl kontrolovat, to co uživatelé sdílejí, má možnost si projít nové klíčové slova a překlady. Jestli je uzná za vhodné, může je přidat do oficiálního slovníku.

7.5.1. Odstranění redundancí

Předtím, než začne samotná kontrola, doporučuje se vyčištění internetových referencí, příkladů a sémantických vztahů, z důvodů ulehčení práce správce. Po tomto pročištění se vymažou data z modifikovaného slovníku, které jsou také obsaženy v oficiálním slovníku. Toto je naimplementováno, protože v případě hodnocení překladu, který je již obsažen v oficiálním slovníku se kopírují veškerá data. Tedy když chce uživatel přidat `user_freq`, tak se zkopírují i internetové reference apod., které jsou v tomto překladu.

```

//vycistení Ex -----
if (modDef.Ex.Count > 0)
{
    List<App_Code.AttributesOne> pom = new
    List<App_Code.AttributesOne>();

    foreach (App_Code.AttributesOne r in modDef.Ex)
    {
        foreach (App_Code.AttributesOne s in ofiDef.Ex)
        {
            if (r.Equals(s))
            {
                pom.Add(s);
            }
        }
    }
    foreach (App_Code.AttributesOne r in pom)
    {
        modDef.Ex.Remove(r);
    }
}

```

Zdrojový kód 8 mazání příkladů

Postup pro mazání redundancí u příkladů je ukázán v Zdrojový kód 8. Příklady se prochází postupem „každý s každým“. Při nalezení shody je tento příklad přidá do seznamu a posléze se tento seznam projde a z modifikovaného slovníku se tyto data vymažou. Obdobně se postupuje i u internetových referencí a sémantických vztahů.

7.5.2. Navigace v slovníku

Veškerá navigace mezi klíčovými slovy a překlady je realizována pomocí inkrementace, popřípadě dekrementace ukazatele na místo v seznamu. Když je ukazatel navýšen nad hodnotu, kterou má určenou jako maximální, tak se nastaví na 0 a přejde se tedy zpět na začátek seznamu. Pro příklad ukázáno na inkrementaci definice (Zdrojový kód 9). Po změně se aktualizují zobrazené záznamy.

```
if (defCur < defMax)
    defCur++;
else
    defCur = 0;
changeModTexts(mod.Def[defCur]);
```

Zdrojový kód 9 Navýšení ukazatele na definici

Správce prochází pouze klíčovými slovy a definicemi modifikovaného slovníku. Když existují i v oficiálním slovníku, tak se zobrazí také, aby byla možnost porovnání.

7.5.3. Přidání překladu

Správce má na výběr několik možností, jak přidat konkrétní překlady z modifikovaného slovníku.

- Klíčové slovo: přidají se všechny definice a atributy klíčového slova.
- Definice: přidá se konkrétní definice a atributy klíčového slova.
- Specifické atributy: přidá se pouze konkrétní atribut.

Klíčové slovo

Tímto způsobem nelze přidat pozměněné atributy. Ty budou vloženy přesně takové, jak jsou uloženy v modifikovaném slovníku. V případě, že oficiální slovník neobsahuje konkrétní klíčové slovo, tak se přidá se celé do oficiálního slovníku (Zdrojový kód 10). Definice se nepřidávají přímo, ale přes pomocnou proměnnou, z důvodu zadávání hodnoty `UserFreq` v modifikovaném slovníku. Protože v oficiálním slovníku nemůže být uživatelské hodnocení, tak je tato hodnota je nastavena na 0.

```
ofi.K = mod.K;
List<App_Code.Definition> pom = new List<App_Code.Definition>();
foreach (App_Code.Definition def in mod.Def)
{
    App_Code.Definition defi = new App_Code.Definition(def);
    defi.UserFreq = 0;
    pom.Add(defi);
}
ofiDic.Lexicon.Add(ofi.K, pom);
ofi.Def = ofiDic.Lexicon[ofi.K];
changeOfficialTexts(ofi.Def[defCur]);
txtOKey.Text = ofi.K;
```

Zdrojový kód 10 Neexistuje klíčové slovo v oficiálním slovníku

Pokud klíčové slovo oficiální slovník obsahuje, tak se prochází definice v obou slovnících způsobem každý s každým. V případě nalezení shody jsou přidány, popřípadě jsou upraveny atributy a je

vypočtena nová hodnota frekvence. Jelikož modifikovaný slovník neobsahuje redundantní data, lze přidat seznamy příkladu atd. V případě, že nová hodnota `del` je větší nebo rovna hodnotě nastavené v konfiguračním souboru pro daný jazyk, tak se definice zneplatní.

```
d.Dtrn = c.Dtrn;
d.Categ = c.Categ;
d.Gr = c.Gr;
d.Etm = c.Etm;
d.Kref = c.Kref;
CheckFreq(); //aktualizace frekvence
if (mod.Def[defCur].Sr.Count() != 0)
    d.Sr.AddRange(mod.Def[defCur].Sr);
if (mod.Def[defCur].Ex.Count() != 0)
    d.Ex.AddRange(mod.Def[defCur].Ex);
if (mod.Def[defCur].Iref.Count() != 0)
    d.Iref.AddRange(mod.Def[defCur].Iref);
found = true;
if (d.Del >= App_Code.GlobalVariables.Config[lang])
{
    d.Invalidate = true;
    temp.Add(d);
}
changeOfficialTexts(ofi.Def[defOCur]);
```

Zdrojový kód 11 Existuje překlad v oficiálním slovníku

V případě, že není nalezen překlad, postupuje se podobným způsobem, jak bylo popsáno výše (Zdrojový kód 11). Rozdílem je však, že místo úprav se tento překlad přidá pod klíčové slovo.

Definice

V definici lze modifikovat data před přidáním do oficiálního slovníku. Logika zde funguje obdobně, jako bylo popsáno výše (Klíčové slovo). Mění se chování podle toho, jestli existuje klíčové slovo, případně jeho definice. Když neexistuje klíčové slovo, tak se přidá s pouze vybranou definicí. V případě známé definice se přidají atributy, které jsou momentálně zobrazeny. Tímto způsobem se zajišťuje případná modifikace záznamů.

Specifické atributy

Použit lze pouze použít v případě, že v oficiálním slovníku existuje překlad, který je v současné chvíli zobrazen. Tento atribut se přidá, popřípadě se změní záznam v oficiálním slovníku (Zdrojový kód 12). Určeno pro případ, kdy správce chce použít pouze část překladu.

```
if (ofi.Def[defOCur].Def == mod.Def[defCur].Def)
{
    ofi.Def[defOCur].Categ = txtCat.Text;
    txtOCat.Text = txtCat.Text;
}
```

Zdrojový kód 12 Změna kategorie

7.6. Komprimace slovníku

Způsob komprese dat byl inspirován návodem na stránkách MSDN ¹ [25] (Microsoft Developer Network). Při každém vytvoření nové verze slovníku je rovněž vytvořena jeho komprimovaná podoba (Zdrojový kód 13).

```
using (FileStream inFile = fi.OpenRead())
{
    // Prevent compressing hidden and already compressed files.
    if ((File.GetAttributes(fi.FullName)
        & FileAttributes.Hidden)
        != FileAttributes.Hidden & fi.Extension != ".gz")
    {
        // Create the compressed file.
        using (FileStream outFile =
            File.Create(fi.FullName + ".gz"))
        {
            using (GZipStream Compress =
                new GZipStream(outFile,
                    CompressionMode.Compress))
            {
                // Copy the source file into the compression stream.
                inFile.CopyTo(Compress);
            }
        }
    }
}
```

Zdrojový kód 13 Komprese dat

Jak již bylo popsáno v analýze (kapitola 5.6), výsledný formát komprimovaného souboru je GZip. Při vytvoření nového komprimovaného souboru se starý soubor přepíše.

¹ *Microsoft MSDN: How to compress files* [online]. 2011 [cit. 2016-03-12].
<https://msdn.microsoft.com/library/ms404280%28v=vs.100%29.aspx>

8. Testování

Výsledný produkt jsem testoval osobně a také jsem jej dal otestovat předpokládané komunitě uživatelů (akceptační testování), tedy mým kolegům na vysoké škole. Jedné skupině testerů byl nabídnut přístup k celému zdrojovému kódu (white box^{1,2} [26][27]). Druhá skupina testovala pouze program, aniž by věděla, jak program funguje (black box [26][27]). Odzkoušeli stránky z pohledu uživatelů a správce slovníků zvlášť. Díky tomu jsem získal užitečnou zpětnou vazbu.

Většina byla spokojena, až na 2 osoby. Ti měli připomínky ke složitosti grafického rozhraní. Toto však není možné upravit, protože stránky jsou již hodně zjednodušené.

Také jsem dal systém odzkoušet „amatérským“ uživatelům. Tedy těm, kteří se aktivně nezabývají informačními technologiemi. Těmto jsem dal pouze k otestování stránky z pohledu uživatelů, protože se od nich největší pravděpodobností nebude očekávat kontrola překladů. Tato skupina testovala pouze black box přístupem.

Tito uživatelé shledali stránky jednoduché a intuitivní. Jedna z připomínek byla na barvu tlačítek stáhnutí překladů. Po provedení změn byli spokojeni.

Mým konzultantem byla také Mgr. Alena Kašpárková, Ph.D., pracující jako profesionální překladatelka. Líbí se jí kategorie překladů, četnost výskytů a možnost rozšiřování, a také vytváření nových slovníků. Poskytla mi mnoho cenných připomínek k slovníku.

Tímto způsobem jsem dovršil své úsilí vytvořit funkční systém. Uvědomuji si, že je pravděpodobné, že se dalším užíváním mohou zjistit další problémy, které bude potřeba vyřešit.

¹ PATTON, R. Testování softwaru. SAMS, Computer Press, Praha, 2002, ISBN 80-7226-636-5.

² PALETA, Petr. *Co programátory ve škole neučí: aneb Softwarové inženýrství v reálné praxi*. Vyd. 1. Brno: Computer Press, 2003, 337 s. ISBN 80-251-0073-1.

9. Budoucí rozšíření

Jako všechny programy je možno i tento dále rozvíjet. Při vytváření tohoto systému jsem se snažil potřeby uživatelů předvídat. I když nelze vyhovět všem, lze předpokládat, že doposud implementovaná kombinace tlustého klienta a webového serveru poslouží jako základ, na kterém se posléze bude moci dále stavět. Jak jsem se již zmínil Mgr. Alena Kašpárková, Ph.D. mi také sdělila několik nápadů, které bych mohl dále implementovat. Po úpravě systému je nutné přistoupit k jeho dalšímu testování.

Možnost vrácení změn

Při kontrole modifikovaného slovníku má správce různé možnosti přijetí klíčových slov a překladů (kapitola 5.5.3). Po přijetí překladu však správce nemůže svoje rozhodnutí vrátit. Správce si tedy musí být jistý svým rozhodnutím přijmout překlad. Bylo by tedy efektivní, aby systém podporoval možnost vrácení změn, alespoň o 3 nazpět.

Lokalizace

V současnosti jsou stránky vytvořené v anglickém jazyce. Pro rozšíření počtu slovníků, tak by bylo dobré vytvořit lokalizace stránek i do dalších jazyků. Ty by se určily dle podporovaných jazyků slovníků.

Data překladu

Formát XDXF nabízí velkou možnost v oblasti obsahu samotného překladu. Odhad, co uživatelská komunita bude potřebovat a používat v překladech jsem vytvořil podle mých zkušeností a potřeb.

Po konzultaci byly zjištěny možné úpravy:

- Odebrat etymologický význam slova. Průměrný uživatel jej nebude potřebovat znát, tedy je možné jej odstranit.
- Přidat redakční komentář, který označuje text, objasňující smysl nebo kontext. Je tedy blíže vysvětlen, význam překladu.
- Přidat přehrávání výslovnosti slova. Nevýhodou je navýšení velikosti slovníku (MB).
- Přidat zkratky.

Tyto úpravy jsou popsány blíže v příloze Formát XDXF.

Uložení jména správce do překladu

Při kontrole slovníku správcem by se ukládalo jméno správce k překladu. To by sloužilo převážně k tomu, že specifický správce bude například lékař. Tudíž by měl na starosti překlady, které jsou medicínského charakteru. Také by vznikla zpětná kontrola, pomocí které by bylo možné zjistit, kdo kontroloval překlad.

Bližší specifikace jazyka

Jazyky mají několik obdob. Například angličtina může být americká, britská, popřípadě australská. V současnosti slovník nerozlišuje jejich verze, tudíž jediný způsob jak je rozlišit, je napsat to do jednoho z možných atributů, jako je například definice.

Podepsání XDXF souboru

Jelikož k souboru může kdokoli přistupovat a měnit jej pomocí jakéhokoli textového editoru, obsah není zabezpečený. Hrozí, že někdo pozmění data způsobem, že budou pro server nečitelná, popřípadě ho zahltní zbytečnými údaji. Tomuto by se dalo předejít, kdyby byl XDXF slovníkový soubor podepsán. Tedy by soubor dokázal upravovat pouze server a klient.

10. Závěr

Cíl zadání byl splněn: Webový server byl navržen a implementován. Bylo tak učiněno s ohledem na existující klientskou část i standardy tagů XML pro slovníky. Implementovaný prototyp byl shledán funkčním, protože byl s dobrovolníky úspěšně proveden experimentální test funkčnosti a rozhraní webového serveru. Celý návrh je formálně zdokumentován. Svůj slovník jsem předal komunitě vývojářů na serveru sourceforge¹ [28] a doufám, že se na jeho vývoji bude pokračovat.

Dovoluji si připojit ještě několik osobních vyjádření:

Tento server jsem se snažil vytvořit co nejlépe a vynaložil jsem k tomuto cíli mnoho úsilí a času. V průběhu práce mě napadaly další a další nápady pro inovaci a rozšíření serveru. Část se mi jich povedlo implementovat, ale vzhledem k omezenému času jsem nevyužil všechny. I když by bylo možné k řešení přistupovat jinou cestou, jsem si jist, že způsob, který jsem si zvolil já, je z hlediska funkčního prototypu nejvhodnější. Při vývoji jsem měl radost, že jsem si vyzkoušel různé technologie a jejich funkčnost. Velmi mě potěšilo, že jsem navázal na mou bakalářskou práci [2] a tím vytvořil způsob, jak sdílet a spravovat volně šiřitelné slovníky.

¹ *SourceForge - User Extensible Dictionary* [online]. [cit. 2016-04-20]. Dostupné z: <https://sourceforge.net/projects/user-extensible-dictionary/>

11. Reference

- [1] Stackoverflow. *C#. Stackoverflow* [online]. 2013 [cit. 2016-03-25]. Dostupné z: <http://stackoverflow.com/questions/tagged/c%23>
- [2] HLOŽÁNKA, Vojtěch *Uživatelsky rozšiřitelný slovník (User-Extensible Dictionary)*. Ostrava, Obhájeno 2014. Bakalářská práce. VŠB - Technická univerzita Ostrava, Fakulta elektrotechniky a informatiky, Katedra informatiky. Vedoucí práce doc. RNDr. Petr Šaloun Ph.D.
- [3] *GNU/FDL Anglicko-Český slovník* [online]. [cit. 2016-03-31]. Dostupné z: <http://slovník.zcu.cz>
- [4] *Eterna Game* [online]. [cit. 2016-03-18]. Dostupné z: <http://www.eternagame.org/web/>
- [5] *European space agency* [online]. [cit. 2016-03-18]. Dostupné z: <http://www.esa.int/gsp/ACT/ai/projects/astrodrone.html>
- [6] *Made* [online]. [cit. 2016-04-06]. Dostupné z: <http://www.made.com/about-us>
- [7] *Duolingo* [online]. 2013 [cit. 2016-03-10]. Dostupné z: <https://www.duolingo.com/>
- [8] What is reCAPTCHA. *ReCaptcha* [online]. [cit. 2015-11-10]. Dostupné z: <https://www.google.com/recaptcha/intro/index.html>
- [9] GOOGLE. *Google translate* [online]. [cit. 2016-02-12]. Dostupné z: <https://support.google.com/translate/#>
- [10] *Seznam* [online]. [cit. 2016-03-20]. Dostupné z: <https://www.seznam.cz/>
- [11] XDXF Description. *Github* [online]. [cit. 2016-02-01]. Dostupné z: https://github.com/soshial/xdxf_makedict/blob/master/format_standard/xdxf_description.md
- [12] GoldenDict [online]. [cit. 2016-02-02]. Dostupné z: <http://goldendict.org/>
- [13] XDClient [online]. [cit. 2016-02-02]. Dostupné z: <http://savannah.nongnu.org/projects/xdclient/>
- [14] MLÝNKOVÁ, Irena, POKORNÝ, Jaroslav. *XML technologie: principy a aplikace v praxi*. 1. vyd. Praha: Grada, 2008, 267 s. Průvodce (Grada). ISBN 978-80-247-2725-7.
- [15] *Source Forge* [online]. [cit. 2016-03-10]. Dostupné z: <https://sourceforge.net/projects/xdxf/>
- [16] *NET Framework System Requirements* [online]. [cit. 2016-03-22]. Dostupné z: <https://msdn.microsoft.com/en-us/library/8z6watww%28v=vs.100%29.aspx>
- [17] FOWLER, Martin. *P of EAA* [online]. 2000 [cit. 2016-02-25]. Dostupné z: <http://martinfowler.com/eaCatalog/dataMapper.html>
- [18] MARTINOVIČ, Jan. *Overview .Net*. Dostupné také z: <http://ita.vsb.cz/11p%C5%99edn%C3%A1%C5%A1ky.aspx>
- [19] *Microsoft MSDN: ASP.Net security Architecture* [online]. [cit. 2016-03-12]. <https://msdn.microsoft.com/en-us/library/yedba920.aspx>
- [20] *Microsoft MSDN: FileUpload Class* [online]. [cit. 2016-03-12]. <https://msdn.microsoft.com/en-us/library/system.web.ui.webcontrols.fileupload%28v=vs.110%29.aspx>
- [21] *Microsoft MSDN : Administration* [online]. [cit. 2016-04-10]. Dostupné z: <https://msdn.microsoft.com/en-us/library/ssa0wsyf.aspx>
- [22] *GZip* [online]. [cit. 2016-03-10]. Dostupné z: <http://www.gzip.org/>
- [23] *Stack overflow*. [online]. 4.5.2012 [cit. 2016-03-04]. Dostupné z: <http://stackoverflow.com/questions/10454519/best-way-to-compare-two-complex-object>

-
- [24] *Net Informations* [online]. [cit. 2016-03-04]. Dostupné z: <http://net-informations.com/faq/net/shallow-deep-copy.htm>
- [25] *Microsoft MSDN: How to compress files* [online]. 2011 [cit. 2016-03-12]. <https://msdn.microsoft.com/library/ms404280%28v=vs.100%29.aspx>
- [26] PATTON, R. Testování softwaru. SAMS, Computer Press, Praha, 2002, ISBN 80-7226-636-5
- [27] PALETA, Petr. *Co programátory ve škole neučí: aneb Softwarové inženýrství v reálné praxi*. Vyd. 1. Brno: Computer Press, 2003, 337 s. ISBN 80-251-0073-1.
- [28] *SourceForge - User Extensible Dictionary* [online]. [cit. 2016-04-20]. Dostupné z: <https://sourceforge.net/projects/user-extensible-dictionary/>

Příloha A. Obsah přiloženého DVD

- Text diplomové práce
- Webový server a jeho zdrojový kód
- Data slovníků
- XSD formátování XDXF

Příloha B. Formát XDXF

Tento standard byl již popsán v mé předchozí bakalářské práci [2]. Každý slovník je umístěn v své vlastní složce a název složky se používá jako ID. Musí obsahovat pouze znaky latinské abecedy a nesmí obsahovat mezery nebo jiné speciální znaky. Takže, pokud jméno slovníku je "Webster's Unabridged Dictionary published in 1913", pak název složky by mohl být např. "Webster1913". Soubor obsahující samotný slovník se pak musí vždy jmenovat "dict.xdxf". Je doporučeno, aby každý slovník měl sadu ikon pro panel nástrojů a velkou ikonu na hlavní straně. Rozměry by měly být: 16×16, 32×32, 512×512. A názvy souborů icon16.png icon32.png a icon512.png. Názvy souborů rozlišují malá a velká písmena.

Všechny XDXF slovníky mají příponu ".xdxf". Jsou ve formátu XML s jakýmkoliv kódováním Unicode (obvykle UTF-8). Jakékoliv jiné kódování je přísně zakázáno.

Tagy

```
<xdxf lang_from="XXX" lang_to="XXX" format="FORMAT" revision="DD">
```

XML 7 Formát hlavičky [11]

Kořenový element, musí mít 4 atributy (XML 7).

- lang_from: 3písmenné slovo, reprezentující z jakého jazyka je slovník.
- lang_to: 3písmenné slovo, reprezentující do jakého jazyka překládáme; oba atributy jsou dle normy ISO 639-3.
- format: Atribut definuje výchozí formátování slovníku a může být buď visual, nebo logical. Výchozí formát může být přepsán pro konkrétní položky (viz níže).
 - Ve formátu visual jsou položky formátovány vizuálně. Jsou určeny pro slovníkové programy (dále uvedený jako SP). Bez vkládání nebo odstraňování mezer nebo EOL (konec řádku). Nicméně, SP mohou označit obsah některých logických tagů (jako <gr> nebo <abbr>) s různými barvami.
Vizuální formát se nedoporučuje! Formát XDXF je vyvinut speciálně pro logicky strukturované slovníky. Vizuální formát je zde poskytnut pouze za tím účelem, aby umožnil kompatibilitu se slovníky, které jsou převedeny ze starých formátů (plain-text).
 - V logickém formátu položky formátovány vizuálně nejsou.
- revision: určuje verzi ve kterém je soubor XDXF naformátován.

Slovník je rozdělen do dvou částí: `<meta_info>` a `<lexicon>` (XML 8) .

```

<xdxf ...>
  <meta_info>
    All meta information about the dictionary: its title, author etc.
  </meta_info>
  <lexicon>
    <ar>article 1</ar>
    <ar>article 2</ar>
    <ar>article 3</ar>
    <ar>article 4</ar>
    ...
  </lexicon>
</xdxf>

```

XML 8 rozložení dokumentu [11]

`<meta_info>` obsahuje veškeré „meta“ informace ohledně slovníku.

1. `<title>` Zkrácený název slovníku napsán v angličtině.
2. `<full_title>` Celý název slovníku. Může obsahovat ne-anglický název.
3. `<publisher>` Oficiální vydavatel slovníku (nepovinný údaj).
4. `<authors>` Obsahuje seznam `<author>` tagů a obsahuje všechny osoby (organizace), které se podílely na tomto slovníku (nepovinný údaj).
 - `<author role="xxx">` Jeden tag pro každého autora.
Jedena osoba může mít více rolí, tudíž by měla být uložena pomocí dvou `<author>` tagů
5. `<description>` Popis slovníku od vydavatele. Doporučuje se zahrnout následující: Autorská práva, licence, odkud tento soubor lze stáhnout, odkud neformátovaný soubor (tj. původní soubor slovníku před konverzí do formátu XDXF) může být stáhnut, odkud původní neformátovaný slovník soubor byl získán, a odkaz na skript, který byl použit k převodu původní neformátovaný soubor slovníku do XDXF. Popis může obsahovat XHTML tagy, které jsou povoleny v XDXF (viz níže).
6. `<abbreviations>` je seznam `<abbr_def>` tagů (XML 9). Obsahuje všechny zkratky používané pro popisky ve slovníku (gramatické popisky např.). `<abbr_def>` tag definuje zkratky a obsahuje dva různé tagy.
 - `<abbr_k>` Zkratka pro klíč zkratky: zkrácený text.
 - `<abbr_v>` Zkratka pro hodnotu: full text, popis zkratky, který se zobrazí, když nad ním podržíme kurzor.

`<abbr_def>` Může obsahovat `type` atribut, který uvádí, jaký typ popisku tato zkratka je:

- `<abbr_def type="grm">` — Uvádí gramatické vlastnosti slova (podstatné jméno, přídavné jméno, atd.).

- `<abbr_def type="stl">` — Stylistické vlastnosti slova (vulgární, archaický, poetický, atd.).
- `<abbr_def type="knl">` — Oblast/doména znalostí (počítače, literatura, gastronomie, typografie atd.).
- `<abbr_def type="aux">` — Jednoduché pomocné slova jako ("např.", "tedy", "zřídka", atd.).
- `<abbr_def type="oth">` — Ostatní.

```

<abbreviations>
<abbr_def type="grm"><abbr_k>n.</abbr_k><abbr_v>noun</abbr_v></abbr_def>
<abbr_def type="knl"><abbr_k>polit.</abbr_k><abbr_v>politics</abbr_v>
</abbr_def>
<abbr_def><abbr_k>Av.</abbr_k><abbr_k>Ave.</abbr_k><abbr_v>Avenue</abbr_v>
>
</abbr_def>
</abbreviations>

```

XML 9 Příklad abbreviations [11]

7. `<file_ver>`, `<creation_date>` jsou povinné informace. `<last_edited_date>`, `<dict_edition>`, `<publishing_date>`, `<dict_src_url>` jsou nepovinné „meta“ informace. Všechna data by měla být formátována jako DD-MM-YYYY. Pokud datum není zcela známo, použijeme nuly: 05-00-2011.

`<lexicon>` obsahuje veškeré `<ar>` (položky) tagy. Skupiny `<ar>` tagů spojují dohromady věci týkající se jedné z klíčových-frází.

Mohou mít volitelný atribut `f`, např. `<ar f="x">`, který může mít hodnotu buď *v* (visual) nebo *l* (logic), a mohou být použity k potlačení výchozího formátování slovníku, které bylo uvedeno v `<xdxf>` tagu.

Následující dva tagy jsou povoleny pouze mezi `<ar></ar>` tagy.

- I. `<k>` Klíčová fráze je jedinečný sled písmen/ideogramů, kterým je tag identifikován a může být nalezen. Může obsahovat více než jednu klíčovou frázi, ale vždy nejméně jednu. Pokud existuje více než jeden `<k>`, SP by měl zobrazit všechny varianty klíčových-frází. Tag `<k>` nesmí být vnořen v jiném `<k>`.
 - `<opt>` Volitelná součást klíčové-fráze. Položka je prohledána pomocí obsahu `<k>` bez obsahu `<opt>`, ale ukazuje se v položce s ním. Tag `<opt>` může být použit pouze mezi `<k></k>` tagy.
- II. `<def>` Označuje celé tělo položky slova, definice, skupiny definic, které spadají do určité kategorie.

Tip. Tyto kategorie mohou být různé části řeči, nebo mají jiný etymologie, například, nebo stejný smysl s různými konotacemi.

`<def>` tagy se mohou vnořovat a obvykle dělají, pokud položka není překlad „1 slovo-na-1 slovo“.

Tag `<def>` musí být uvnitř `<ar>`, i když položka je jednoduchá a není třeba vytvořit skupinu.

* Mohou mít `cmt` (komentář) atribut, který pomáhá rozlišit jednu definici od ostatních.

* Mohou mít unikátní tribut `id` [01-9a-zA-Z], na který může být odkaz v jiné položce.

* Mohou mít integer/double atribut `freq` (frekvence): absolutní nebo relativní četnost hodnoty definice.

V položkách s tagy `visual` formátu `<def>` nemají vliv na formátování. Pro položky, které mají logický formát SP musí rozlišovat vizuálně jednu definici od druhé podle úrovně vnoření pomocí odsazení, zmenšení velikosti písma nebo číslováním definic '1) ', '2) "... nebo 1. ', '2. "... nebo "A.", "B." ... atd. před každou definicí.

1. `<gr>` Určuje gramatické informace o slově. Mohou obsahovat různé tvary slova, použití slova, gramatické popisky a další informace o tohoto druhu.
2. `<tr>` Označuje přepis/výslovnost informace, IPA symboly jsou výchozí. Může mít také "mode" atribut s hodnotami `X - SAMPA` nebo `erkIPA`.
3. `<kref>` Jednoduchý odkaz na jiný klíčovou frázi, která se nachází ve stejném souboru. Pro další informace viz `<sr>`.
4. `<dtrn>` Tento tag označuje přímý překlad klíčové fráze (obvykle nepoužívané pro vysvětlující slovníky). V ne-výkladových slovnících by měl být tag použit pro pomoc softwaru, aby slovník mohl automaticky extrahovat základní a nejjednodušší překlady klíčové fráze. Může to být užitečné pro:
 - Automatická extrakce dat pro nápovědné překlady (např. jako `qDictionary`).
 - Zviditelnit `<dtrn>` v seznamu slov, aby se zabránilo příliš častému vyhledávání „celé položky“.

Tato slova by měla být vyznačena pro jednodušší prohledávání napříč položkami (obvykle tato slova jsou zobrazena jako tučné, někdy také v tmavě - oranžové). Tato slova mohou také automaticky vypadat jako odkazy (`<kref>`), které vedou k položce opačného jazykového páru.
5. `<rref>` Odkaz na zdroj audio souborů (`mp3`, `ogg` nebo `opus` formát), který by měl být umístěn ve stejné složce jako slovník (XML 10).
 - volitelné `start` a `size` atributy jsou nezbytné pro audio a video soubory, kdy referenční body ukazují na určité části souboru. To je velmi výhodné, abychom měli všechny audio data slovníku v jednom souboru, s odkazem na různé části z různých překladů. Atribut `start` určuje offset: pozici v souboru prvního bajtu bloku zájmu, a `size` určuje jeho délku v bajtech. Je-li `start` atribut vynechán, pak se předpokládá, že to je 0. Je-li atribut `size` vynechán, pak se předpokládá, že soubor má být přehrán až do konce.

```
<rref start="xxx" size="xxx">crawl</rref> nebo <rref
start="xxx" size="xxx"/>
```

6. `<iref href="http://www.somewebsite.com">` Odkaz na zdroj v Internetu. Zachování předponu („http://“, „https://“), je povinné.
7. `<abbr>` Označuje zkrácený tag. Tagy by měly mít vysvětlení uvedeny v korespondenčním `<abbr_k>` tagu v `<abbreviations>` části.
8. `<c c="xxxxxx">...</c>` (označuje 32 bitový (6-místný) hexadecimální kód barvy) Označuje text danou barvou. Syntaxe pro "c" atribut je stejný jako pro "color" Atribut "font" tagu v HTML. Pokud je atribut barva vynechán, bude vybraná výchozí barva. Výchozí barva je vybraná SP.
9. `<ex>` Označí text příkladu (obvykle uvedený v šedé nebo jiné barvě podle SP). Tip (indexování): příklady, musí být indexovány implicitně, ale uživatelé by měli být schopni nastavit, zda chtějí příklady, které budou indexovány a vyhledávány, nebo ne. Atribut `type`, může být:

- `exm` - obvyklé příklady s nebo bez překladů
- `phr` - může obsahovat jakýkoli typ frazémů (idiomy, kolokace, klišé atd.)
- `prv` - přísloví
- `oth` - jiné

Atributy `source` a `author` určují, odkud byl příklad převzat.

Měl obsahovat tagy: `<ex_orig>` pro původní věty příkladu (v hodnotě: 1 nebo více).

`<ex_tran>` Je volitelný, může být několik překladů (počet: 0 nebo více).

`<mrkd>` Je volitelný tag, který se používá k označení se cílového slovo v původní frázi a překladu

10. `<co>` Označuje text redakčního komentáře, který objasňuje smysl nebo kontext (zobrazeno v jiné barvě programem, obvykle šedá).
Tip (indexování): komentáře jsou obvykle indexovány.

11. `<sr>` Je sekce věnovaná sémantickým vztahům k jiným slovům, jako jsou synonyma, holonyma, hyperonyma apod (XML 11). Využívá `<kref>` k adresování ostatních slov s dodatečnými atributy tagů.

`<sr>` část vypadá takto:

```
<sr>
<kref type="syn" kcmt="obsolete">game</kref>
<kref type="hol" kcmt="partly">play</kref>
</sr>
```

XML 11 příklad sémantických vztahů [11]

Možné `type` hodnoty:

- `syn` a `ant` - synonyma a antonyma;
- `hpr` a `hpn` - hyperonyma a hyponyma (začleňují toponyma);
- `par` a `spv` - paronyma a pravopisné varianty;
- `mer` a `hol` - meronyma a holonyma;

- `ent` - vyplývání, speciální sloveso kategorie: v. Y je spojeno X, pokud používáte X, musíte použít Y, může být také použito pro podstatná jména: podstatné jméno s sebou nese nějakou jinou věc (dělá "zločin" s sebou nese, že je "zločinec");
- `rel` - označuje význam (např., „hezké“ a „ošklivé“, jsou důležité pro „vzhled“);

`kcmt` atribut se používá k určení, do jaké míry je toto slovo souvisí s naším nebo specifikovat další informace o slově (jako je pohlaví nebo případ)

12. `<etm>` etymologické informace o slově.
13. `<di>` Označuje část `<def>` textu, které by neměly být indexovány. Mohly by být použity pouze v `<def>` a některé její děti: `<co>`, `<ex>`, `<etm>`, `<phr>`.
14. `<categ>` tag je poněkud ekvivalent kategorií Wikipedii (např. „výslech“ by mohlo mít kategorii „Středověká historie“)